ARO Report 88-2

PROCEEDINGS OF THE THIRTY-THIRD CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH DEVELOPMENT AND TESTING

Approved for public release; distribution unlimited.
The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Sponsored by
The Army Mathematics Steering Committee
on Behalf of

THE CHIEF OF RESEARCH, DEVELOPMENT AND ACQUISITION

U. S. Army Research Office

Report No. 88-2

May 1988

PROCEEDINGS OF THE THIRTY-THIRD CONFERENCE ON THE DESIGN OF EXPERIMENTS

Sponsored by the Army Mathematics Steering Committee

HOST

The Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

21-23 October 1987

HELD AT

The University of Delaware
Newark, Delaware

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, North Carolina

FOREWORD

The Thirty-Third Conference on the Design of Experiments in Army Research, Development, and Testing was held 21-23 October 1987 on the campus of the University of Delaware. This university served as one of its hosts, the other host being the Ballistic Research Laboratory (BRL). Professor Henry B. Tingey was the Chairperson on Local Arrangements for the University and Dr. Malcolm Taylor served in this capacity of BRL. The members of the Army Mathematics Steering Committee (AMSC), sponsors of these conferences, would like to take this opportunity to thank these gentlemen for their excellent handling of the many problems associated with a meeting of this size.

Members of the Program Committee for the conference were pleased to obtain the services of the following invited speakers to talk on topics of interest to Army personnel:

Speaker and Affiliation	Title of Address
Dr. J. Stuart Hunter Private Consultant	Statistics and the Learning Process
Professor Albert Paulson Rensselaer Polytechnic Institute	A Generalized Likelihood Approach to Experimental Design, Data Analysis and Modeling
Dr. William A. Gale Bell Communications Research	Structural Statistical Knowledge for Expert Systems
Professor Howard M. Taylor University of Delaware	The Effect of Size on Material Strength

On 19-20 October 1987, two days before the start of the Design Conference, a tutorial entitled "Regression Diagnostics" was held. Its speaker was Professor Roy Welsch of the Massachusetts Institute of Technology, Cambridge, MA. The main purpose of these seminars was to develop, in Army scientists, an interest in and and appreciation for the statistical methods that are needed to analyze experimental data.

Dr. J. Stuart Hunter, Professor Emeritus of Princeton University, was the recipient of the seventh Wilks Award for contributions to Statistical Methodologies in Army Research, Development, and Testing. This honor was bestowed on Dr. Hunter for his many significant contributions to various fields of statistics, in particular to the areas of fractional factorial and response surface experimental design. He has assisted many Army scientists with their statistical problems, and has been an invited speaker at four of these Design conferences.

The AMSC has requested that these transactions be published and distributed Army-wide so that the information in them might assist Army scientists with some of their statistical problems. Committee members would like to thank all the speakers for their interesting presentations and also the members of the Program Committee for their many contributions to this scientific meeting.

PROGRAM COMMITTEE

Carl Bates	David Cruess	Eugene Dutoit
Robert Launer	Carl Russell	Douglas Tang
Malcolm Taylor	Jerry Thomas	Henry Tingey

TABLE OF CONTENTS*

*This Table of contents contains only the papers that are published in this tehcnical manual. For a list of all papers presented at the Thirty-Third Conference on the Design of Experiments, see the Program of this meeting.

AGENDA

THIRTY-THIRD CONFERENCE ON THE DESIGN OF EXPERIMENTS

IN ARMY RESEARCH, DEVELOPMENT AND TESTING

21-23 October 1987

Hosts: The Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

and

The Department of Mathematical Sciences
The University of Delaware
Newark, Delaware

Location: Pencader Hall, Room 106
The University of Delaware

* * * * * Wednesday, 21 October * * * * *

0815-0915 REGISTRATION - Clayton Hall Lobby

0915-0930 CALL TO ORDER - Pencader Hall, Room 106

Dr. Malcolm Taylor, Ballistic Research Laboratory

OPENING REMARKS

Dr John T Frasier
Director, Ballistic Research Laboratory

WELCOMING REMARKS

Dr Ivar Stakgold
Chairman, Department of Mathematical Sciences
The University of Delaware

0930-1200 GENERAL SESSION I

Chairman: Prof Henry B Tingey, University of Delaware

0930-1030 KEYNOTE ADDRESS

J Stuart Hunter, Princeton, NJ

1030-1100 BREAK

1100-1200 A BAYESIAN APPROACH TO THE DESIGN AND ANALYSIS OF COMPUTATIONAL EXPERIMENTS

Toby J Mitchell* and Max Morris, Oak Ridge National Labs

1200-1330 LUNCH

1330-1700 CLINICAL SESSION A

Chairman: Barry Bodt, Ballistic Research Laboratory

Panelists: Prof John Green
Prof Vincent LaRiccia
Prof John Schuenemeyer
Prof Robert Stark
Prof Howard Taylor
The Department of Mathematical Sciences
The University of Delaware

ANALYSIS OF A REPEATED DESIGN WITH MISSING CELLS

Michelle R Sams and Joel H Fernandez, White Sands Missile Range

ALTERNATIVE METHODS FOR RELIABILITY ESTIMATION

Raymond V Spring, US Army Natick R&D Directorate
Thomas A Mazzuchi, The George Washington University

ALLOCATION AND DISTRIBUTION OF 155 MM HOWITZER FIRE

Ann E M Brodeen and Wendy A Winner,
The Ballistic Research Laboratory

1500-1530 Break (as needed)

A SIMPLE MATHEMATICAL MODEL FOR THE SIMULATION OF IR BACKGROUNDS

Denis F Strenzwilk, Ballistic Research Laboratory
Walter T Federer and Michael T Meredith, Cornell University

1530-1700 CLINICAL SESSION A, CONTINUED (as needed)

1830-1930 CASH BAR - THE SHERATON INN, NEWARK

1930-2130 BANQUET AND PRESENTATION OF WILKS AWARD - THE SHERATON INN

* * * * * Thursday, 22 October * * * * *

0830-1000 TECHNICAL SESSION 1 - STATISTICAL APPLICATIONS

Chairman: Dr Francis Dressel, US Army Research Office

EVALUATION OF CAMOUFLAGE PAINT GLOSS VERSUS DETECTION RANGE

George Anitole and Ronald L Johnson, US Army Belvoir Research,
Development and Engineering Center
Christopher J Neubert, US Army Materiel Command

A 2-STAGE EXPERIMENTAL DESIGN FOR TESTING LARGE SCALE SIMULATIONS

Aqeel A Kahn, US Army Concepts Analysis Agency

BLACK BRANT HAZARD ANALYSIS

Weston C Wolff, White Sands Missile Range

USING A PERSONAL COMPUTER IN STATISTICAL PLANNING AND ANALYSIS

Carl Russell, Army Operational Test and Evaluation Agency

1000-1030 BREAK

1030-1200 TECHNICAL SESSION 2, EXPERIMENT DESIGN AND LINEAR MODELS

Chairman: William Baker, Ballistic Research Laboratory

ONE SIDED TOLERANCE LIMITS FOR RANDOM EFFECTS MODELS

Mark Vangel, US Army Material Testing Laboratory

ESTIMATION OF VARIANCE COMPONENTS AND MODEL-BASED DIAGNOSTICS IN A REPEATED MEASURES DESIGN

Jock O Grynovicki, US Army Human Engineering Laboratory, APG
J W Green, The University of Delaware

MODEL BASED DIAGNOSTICS FOR VARIANCE COMPONENTS IN A GENERAL MIXED LINEAR MODEL

John W Green, The University of Delaware
R R Hocking, The Texas A&M University

CHANGE-POINT REGRESSION WITH UNKNOWN CHANGE POINTS

Robert L. Launer, US Army Research Office

1200-1330 Lunch

1330-1500 TECHNICAL SESSION 3 - STOCHASTIC PROCESSES

Chairman: Dr Eugene Dutoit, US Army Infantry School

SEMIREGENERATIVE PHENOMENA

N U Prabhu, Cornell University

k-LAPLACE PROCESSES

Lee S Dewals, The US Military Academy
Peter A W Lewis, Naval Postgraduate School
Ed McKenzie, University of Strathclyde, Glascow, Scotland

THEORY OF RANDOM MAPPINGS

Bernard Harris, University of Wisconsin - Madison

1500-1530 BREAK

1530-1730 GENERAL SESSION II

Chairman: Dr Malcolm S Taylor, Ballistic Research Laboratory

A GENERALIZED LIKELIHOOD APPROACH TO EXPERIMENTAL DESIGN, DATA ANALYSIS AND MODELING

Albert Paulson, Rensselaer Polytechnic Institute

STRUCTURING STATISTICAL KNOWLEDGE FOR EXPERT SYSTEMS

William A Gale, Bell Communications Research

* * * * * Friday, 23 October * * * * *

0830-1000 TECHNICAL SESSION 4 - STATISTICAL INFERENCE

Chairman: Linda Moss, Ballistic Research Laboratory

ON THE USE OF FACTOR ANALYSIS AS A PREDICTION TOOL

Oskar M Essenwanger, US Army Missile Command

CONSISTENCY OF THE P-VALUE AND A SET OF Q-VALUES IN A SCORING ACCURACY ANALYSIS

Paul Thrasher, White Sands Missile Range

A BAYESIAN METHOD FOR PROJECTING A TOLERANCE LIMIT

Donald Neal and John Reardon, US Army Material Testing Laboratory

COVERING PROBABILITY PROPERTIES OF COMPETING CONFIDENCE INTERVAL METHODS FOR THE RISK RATIO

Craig Morrissette* and Douglas B Tang, Walter Reed Army Institute of Research

1030-1045 BREAK

1045-1200 GENERAL SESSION III

Chairman: Dr Douglas B Tang, Walter Reed Army Institute of Research
Chairman of the AMSC Subcommittee on Probability and Statistics

1045-1100 OPEN MEETING OF THE STATISTICS AND PROBABILITY SUBCOMMITTEE OF THE ARMY MATHEMATICS STEERING COMMITTEE

1100-1200 THE EFFECT OF SIZE ON MATERIAL STRENGTH

Howard M Taylor, University of Delaware

ADJOURN

ANALYSIS OF A REPEATED MEASURES DESIGN WITH MISSING DATA

Michelle R. Sams and Joel H. Fernandez
U.S. Army Materiel Test and Evaluation/
Engineering and Analysis RAM Division
U.S. Army White Sands Missile Range, NM 88002-5175

ABSTRACT

Electronic Maintenance Publication System (EMPS) is a U.S. Army Materiel Command (USAMC) initiative to determine the feasibility of using current technology to electronically display and deliver the contents of Department of the Army Technical Manuals (DATMs) to the maintenance site. The Army Materiel Test and Evaluation Directorate (ARMTE) was tasked to conduct a "side-by-side" comparison of EMPS vs. DATMs and to conduct a human factors evaluation of the EMPS hardware and software. ARMTE conducted the comparison study on the Patriot System at Ft. Bliss, TX from 6 April to 15 May 1987. Ten operator/maintainers (MOS 24T) were trained to use EMPS and then participated in the test phase performing maintenance tasks on the Radar Set (RS) and on the Engagement Control Station (ECS). A 2 x 2 x 7 within-subjects factorial design was planned, with 2 mediums (EMPS, DATMs) performed on 2 major end items (RS, ECS) for 7 types of maintenance tasks. Due to software constraints and Patriot peculiar problems, only 8 of the 28 possible treatment conditions have observations from all the subjects and 2 of the treatment conditions have no observations. Various data estimation procedures were considered and then rejected on the basis of excessive and systematic missing data. Two analyses of variance were conducted on a subset of the original data, which contained the least amount of missing data and were determined to be representative of the maintenance actions. No significant difference was found for the variables of interest (those involving EMPS and DATMs). Based on the results of this study, it was concluded that there is no evidence to suggest that there is any significant difference in time to perform a fault isolation or remove and install task on the PATRIOT system utilizing either EMPS or DATMs. An electronic delivery of maintenance information (as tested in EMPS) appears to be as effective as the traditional medium of paper technical manuals (DATMs).

Comments and suggestions by the panelists and attendees at the conference were greatly appreciated. We are especially indebted to N. Scott Urquhart of New Mexico State University for his guidance throughout the completion of the data analysis.

INTRODUCTION

Maintainability is a major element of system effectiveness. As such, the delivery of maintenance information is a crucial component in the man-machine system. The current delivery medium is through paper technical manuals (DATMs). Many problems have been noted with the paper manuals (e.g., the large number of bulky manuals needed to contain all the information and difficulties encountered keeping the manuals updated and current, the difficulty using the manuals especially in inclement weather, etc.) An alternative delivery medium was sought and tested in the form of Electronic Maintenance Publication System (EMPS). As part of a larger evaluation of EMPS, the Army Materiel Test and Evaluation Directorate at White Sands Missile Range was tasked to conduct a performance ("side-by-side") comparison of EMPS vs. DATMs and to conduct a human factors evaluation. The performance evaluation was based on the speed and accuracy of maintenance actions for the two mediums and is presented in this paper.

METHOD

Subject and Team Selection

A total of ten operator/maintainers (all trained to the T5 PFAS level) were allotted for the study on the basis of availability. Maintenance tasks are normally performed in maintenance teams consisting of a "reader" and a "doer". For the purposes of this study, the ten subjects were divided into two groups on the basis of their GT scores (an index of general intelligence and ability). Five teams were then formed out of each group (each subject participated in two teams). Each team from

Group A was then matched with a team from Group B with approximately the same GT level. This matching was done in order to reduce some of the variance due to the subjects, especially since there was such a small number of subjects in the experiment.

Experimental Design

A 2 x 2 x 7 within-subjects factorial design was planned, with 2 mediums (EMPS, DATMs) performed on 2 major end items (RS, ECS) for 7 types of maintenance tasks. The design was within-subjects in that all teams would participate under all treatment combinations. However, due to the concern of possible asymmetrical transfer effects, a particular team did not participate in the same task twice. For example, when a team performed a particular task utilizing EMPS, a different team matched for general ability performed the same task utilizing paper DATMs.

Task Selection

With the assistance of subject matter experts, it was determined that there were seven types of maintenance actions performed on the RS and ECS. These task types consisted of fault isolation (FI), remove and install (RI), repair and verify (RV), combined tasks (CO) which included FI, RI, and RV times, preventive maintenance checks and services (PMCS), operations (OP), and repair parts and special tools list (RPSTL). The selection of the specific tasks to be performed was influenced by several factors; software capability, the tasks had to be representative of normal maintenance actions, and the concern of face validity.

Training Session

Ten operator/maintainers were familiarized with EMPS in the classroom and given support documentation. They then participated in an on-site training session in their assigned teams. A total of 63 maintenance tasks on the RS and ECS utilizing both EMPS and DATMs were completed in this session.

Testing Session

The teams then participated in the test phase performing a total of 302 separate maintenance actions consisting of the seven types of maintenance actions on the RS and ECS utilizing both EMPS and DATMs.

Data Collection

The errors committed and the total time to complete a maintenance action were recorded by a data collector for each task. A particular data collector would record data for the same task, performed once by a team utilizing EMPS and again by a matched team utilizing DATMs. This was done to reduce variation in the time and error measurements recorded among the data collectors.

Reduction of the Full Factorial Design

Each team was to participate in an equal number of tasks utilizing the two mediums on both major end items for all task types. Halfway through the test phase, it became obvious that due to equipment failure and frequent removal of the subjects for field training exercises, that the full factorial would not be completed as originally planned. Even though generalizability of the results to all types of maintenance actions was a concern, it was determined that those tasks which best utilized the DATMs and EMPS would be an accurate indicator of the efficiency and

feasibility of the mediums.

Through discussions with subject matter experts and the participating subjects, it was determined that two types of tasks best utilized the two mediums. These were fault isolation (FI) and remove and install (RI). These tasks were complex enough to compel the maintainer to actually read and refer to the maintenance material. The other tasks were simple and routine, so that close attention to either medium was not necessary (although they were instructed to actually read and use both mediums in any circumstance). Within the remaining test phase time, the test schedule was revised to include more of the FI and RI type tasks. As a result, there was a large amount of missing data in the other four types of tasks. The seventh task (RPSTL) was conducted only on the ECS, due to software problems, and is not reported here.

RESULTS AND DATA ANALYSIS

A summary of the data collected for maintenance action times is presented in Table 1 and a means bar chart is presented in Figure 1. There are 81 missing observations out of a total of 240. Estimating the missing data would allow investigation of 3-way interactions (type of task x item x medium) and allow generalization to all types of tasks tested. Various data estimation procedures were investigated, with employing stepwise regression for each missing value on the available variables appearing as the most appropriate method (Frane, 1976).

Frane (1976) cautions that the methods for estimating missing data for multivariate analysis depend on several assumptions:

ENGAGEMENT CONTROL STATION

	EMPS						Paper DATMs						
Team[a]	CO	FI	OP	PM	RI	RV	CO	FI	OP	PM	RI	RV	Team[a]
1–2	5.93[1]	4.87[2]	*	*	15.08[4]	.28[1]	6.52[1]	6.62[2]	*	*	12.70[4]	.22[1]	6–7
2–3	21.05[1]	9.15[2]	3.57[1]	*	9.72[3]	.72[1]	24.25[1]	9.17[2]	3.13[1]	*	11.47[3]	.48[1]	7–8
3–4	*	4.17[1]	2.00[1]	3.45[1]	*	*	*	6.30[1]	2.13[1]	3.17[1]	*	*	8–9
4–5	12.52[1]	4.74[2]	*	10.45[1]	19.69[2]	.15[1]	13.40[1]	6.40[2]	*	12.78[1]	20.52[2]	.10[1]	9–10
5–1	*	15.00[1]	*	9.08[1]	50.87[1]	*	*	8.50[1]	*	2.22[1]	54.88[1]	*	10–6
6–7	11.97[1]	3.24[3]	*	6.51[2]	8.63[1]	.30[1]	11.68[1]	3.65[3]	*	5.19[2]	6.42[1]	.28[1]	1–2
7–8	*	6.82[1]	*	*	*	*	*	9.70[1]	*	*	*	*	2–3
8–9	7.35[1]	4.08[2]	7.88[1]	8.68[1]	9.76[2]	.80[1]	11.78[1]	5.38[2]	6.92[1]	8.42[1]	12.04[3]	.68[1]	3–4
10–6	20.60[1]	4.32[2]	7.87[2]	*	15.69[3]	.53[1]	14.03[1]	3.76[2]	5.46[2]	*	12.44[3]	.30[1]	5–1
9–10	*	4.25[1]	*	*	21.88[3]	*	*	3.02[1]	*	*	23.92[3]	*	4–5

Table 1. Reduced time data (in min) collected from six types of maintenance actions on two major end items utilizing EMPS and paper DATMs. The number in the upper right-hand corner of each cell indicates the the number of observations per cell.

Table 1. (continued)

	RADAR SET												
	EMPS						Paper DATMs						
Team[a]	CO	FI	OP	PM	RI	RV	CO	FI	OP	PM	RI	RV	Team[a]
1–2	18.87[1]	1.44[2]	4.93[1]	*	40.36[2]	1.50[1]	33.22[1]	1.59[2]	2.15[1]	*	56.90[2]	1.55[1]	6–7
2–3	10.25[1]	2.88[3]	*	1.33[1]	9.34[2]	.45[1]	11.17[1]	3.12[3]	*	2.07[1]	11.56[2]	.63[1]	7–8
3–4	*	2.05[2]	*	1.95[1]	66.27[1]	*	*	2.64[2]	*	2.65[1]	85.27[1]	*	8–9
4–5	7.52[1]	3.81[3]	*	*	12.89[3]	.60[1]	10.75[1]	5.02[3]	*	*	8.93[3]	.77[1]	9–10
5–1	14.92[1]	4.00[2]	.62[1]	*	9.53[1]	.80[1]	12.28[1]	2.51[2]	.48[1]	*	9.63[1]	1.22[1]	10–6
6–7	14.03[1]	4.23[4]	*	*	10.56[3]	.67[1]	14.08[1]	5.28[4]	*	*	8.96[3]	1.53[1]	1–2
7–8	*	3.84[2]	*	*	16.72[2]	*	*	2.70[2]	*	*	17.75[2]	*	2–3
8–9	16.44[2]	3.08[4]	*	*	25.30[4]	1.14[2]	18.23[2]	3.43[4]	*	*	20.01[4]	1.28[2]	3–4
10–6	12.12[1]	2.68[4]	*	2.33[1]	31.73[2]	.58[1]	11.17[1]	3.09[4]	*	2.47[1]	37.04[2]	.67[1]	5–1
9–10	*	2.99[2]	*	3.20[1]	13.43[1]	*	*	4.18[2]	*	3.52[1]	14.87[1]	*	4–5

* No data was collected for this cell.

[a] Group A consisted of subjects 1-5, Group B consisted of subjects 6-10. Subjects were assigned to two teams within each group (i.e., subject #1 participated in teams "1-2" and "5-1").

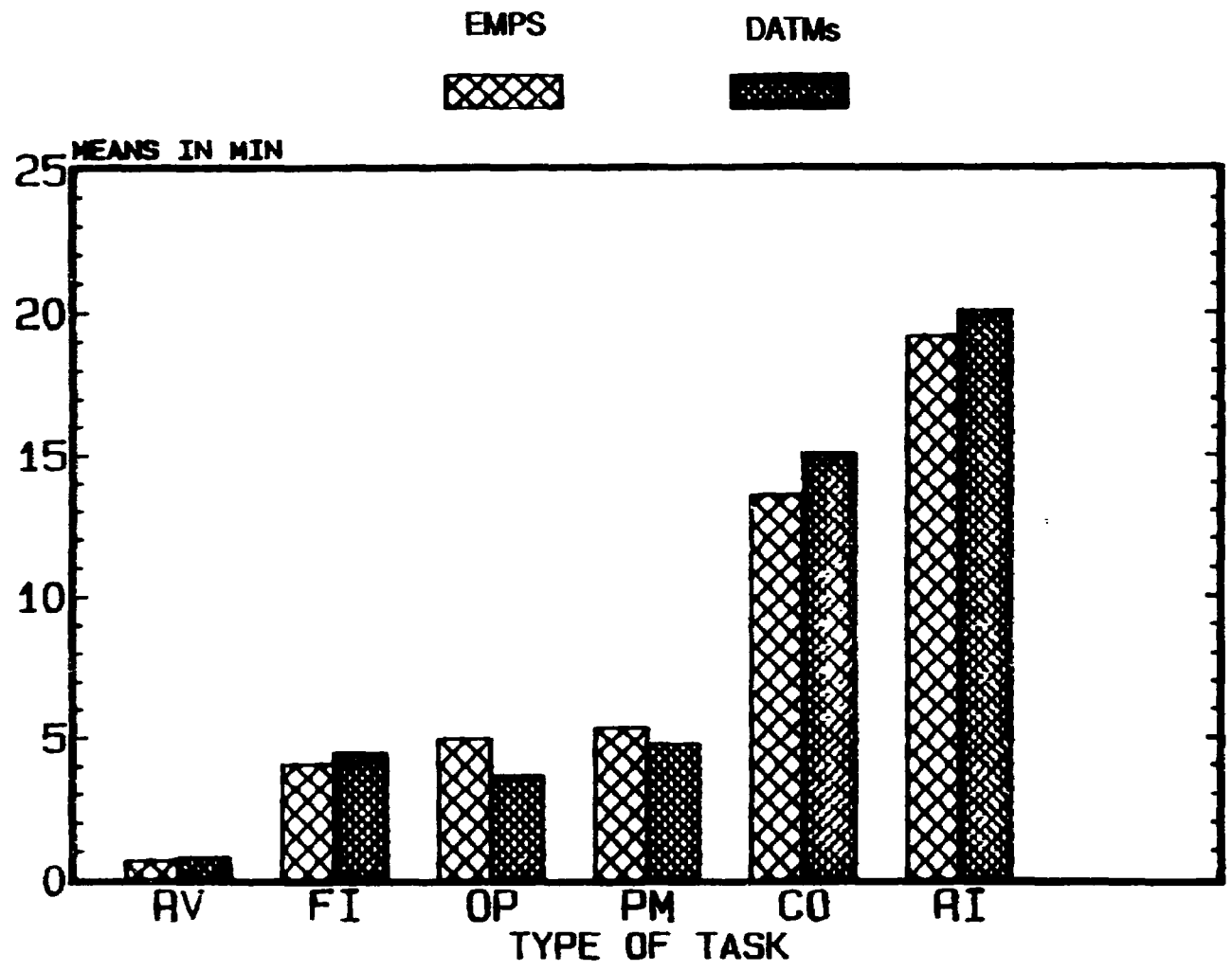


Figure 1. Comparison of mean times collected from six types of maintenance actions utilizing EMPS and DATMs.

data must be missing at random to get a good estimate of the covariance matrix, each missing variable must be highly correlated with one or more available variables, and the amount of missing data should not be excessive. If any of the assumptions are seriously violated, any procedure for handling missing data is likely to be unsatisfactory. The data collected in the study violated two of these assumptions; the missing data was excessive and systematic.

It was determined not to make estimations of the missing observations and to instead conduct separate analyses of variance on the two types of tasks (FI and RI) which contained the least amount of missing data, and which were previously determined to best test the variables of interest. Since the data approximated a lognormal distribution, the data was transformed ((log (X + 1)) to normalize the distribution (Winer, 1971). The transformed maintenance times were subjected to the analyses of variance presented in Tables 2 and 3.

A significant difference for maintenance time for the different tasks within each item was found, ($p < .01$), for both types of tasks. This was neither surprising, nor of interest. The set of tasks performed on each item varied in difficulty. For fault isolation tasks a significant difference was found for item, ($p < .01$). It took longer to perform fault isolation tasks on the ECS than on the RS. Again this was not a variable of interest, and most likely eflects the relative complexity of the the items.

The variables of interest, those involving the two mediums being compared (EMPS and DATMs) revealed no significant differences in maintenance time ($p > .10$). Also there was no

Table 2

ANOVA Table for Fault Isolation Tasks using Log Transformed Time

Source	df	MS	F
Between Subject	43		
Group	1	.007	0.21
Task(Item)	21	.497	14.62 **
Error (Between)	21	.034	
Within Subject	46		
Medium	1	.083	0.82
Item	1	2.930	29.01 **
Medium X Item	1	.003	0.03
Error (Within)	43	0.101	
TOTAL	89		

** $p < .01$

Table 3

ANOVA Table for Remove and Install Tasks using Log Transformed Time

Source	df	MS	F
Between Subject	45		
Group	1	.046	1.24
Task(Item)	22	1.661	44.42 **
Error (Between)	22	.037	
Within Subject	36		
Medium	1	.044	0.73
Item	1	.012	0.20
Medium X Item	1	.025	0.42
Error (Within)	33	0.060	
TOTAL	81		

** $p < .01$

significant difference in group performance for either type of task. Thus the various teams composed from each group were matched fairly well on ability to perform the tasks.

Errors committed while performing the maintenance tasks were negligible and were not subjected to statistical analysis.

CONCLUSION

Based on the results of this study, there is no evidence to suggest that there is any significant difference in time to perform fault isolation and remove and install maintenance actions on the PATRIOT system utilizing either EMPS or DATMs. Errors made while using either medium were negligible and are not a significant factor either. An electronic delivery of maintenance information (as tested in EMPS) appears to be as effective as the traditional medium of paper technical manuals (DATMs).

These are encouraging results considering that the test subjects had a very "quick and dirty" training period with the EMPS system. It is conceivable that the speed with which a maintenance action can be performed with an electronic delivery of maintenance information will improve with a more comprehensive training approach and with Human Engineering improvements to the system.

REFERENCES

Frane, J.W. (1976). Some simple procedures for handling missing data in multivariate analysis. Psychometrika, 41, 409-415.

Winer, B.J. (1962). Statistical principles in experimental design. New York: McGraw-Hill.

ALLOCATION AND DISTRIBUTION OF 155MM HOWITZER FIRE

Ann E. M. Brodeen
Wendy A. Winner

Director,
U. S. Army Ballistic Research Laboratory
ATTN: SLCBR-SE-P
Aberdeen Proving Ground, MD 21005-5066
(301) 278-6659, AV 298-6659

Abstract

The U.S. Army Ballistic Research Laboratory (BRL), Aberdeen Proving Ground, MD, has been investigating the problems associated with allocating and distributing friendly fire based on the importance of an enemy target and its function in a particular tactical situation. The available data contain nonstandard data structures, numerous variables with various degrees of influence on the predictive relationship, a mixture of data types, and nonhomogeneous variable relationships. Various approaches including parametric and nonparametric procedures have been applied to this problem. As an alternative to standard parametric procedures, the BRL is investigating recently published classification tree methodology which extends previous developments in this area [1]. Similar to other classification tree methodologies, this methodology provides predictions by constructing binary trees. However, unlike other analytical techniques, e.g., cluster analysis, linear discriminant analysis, and earlier classification trees, Breiman et al.'s classification tree structured methods concurrently handle these problems, which are common to the data collected by the BRL on Fire Direction Officers' decisions on 155mm howitzer targets.

The authors would like to solicit critiques of the proposed approach to this problem and suggestions for alternatives.

I. Introduction

The U.S. Army Ballistic Research Laboratory (BRL) has been examining the problems associated with selecting the type, volume, and the method of firing ammunition on enemy targets by a specific 155mm howitzer firing configuration, i.e., the allocation and distribution of friendly fire. This research is concentrating on allocating and distributing the fire of 155mm howitzer firing units based on the importance of an enemy target and its function in a particular tactical situation. Results from this research will be incorporated into the BRL's prototype decision aid FireAdvisor. As a tool for developing and implementing fire support plans, FireAdvisor is incorporating commander's criteria, munition effects, and the tactical situation (including firing units, munitions, fuzes, and targets) to assist with determining the optimum allocation and distribution of fire against independent targets over time.

To acquire data for this research, the BRL conducted a statistically designed experiment, the Firepower Control Experiment, in December 1985. In addition, the BRL has recently extracted similar information from scenarios developed by LB&M Associates, Inc., Lawton, OK, under a BRL contract. Both of these data sets are characterized by a mixture of data types, nonhomogeneous variable relationships, and different degrees of influence of the variables. Various approaches such as multiple regression analysis, the Mann-Whitney test, Kruskal-Wallis analysis of variance by ranks, and cluster analysis have been applied to analyze the data from the Firepower Control Experiment. The goals of these procedures were to uncover the relationships among the variables and provide accurate predictions for allocating and distributing 155mm howitzer fire.

As an alternative to standard parametric procedures, the BRL is investigating employing a recently published classification tree methodology to these data sets [1]. Similar to other published classification tree methodologies, Breiman et al.'s methodology provides predictions by constructing binary trees. However, unlike other analytical techniques, Breiman et al.'s classification tree structured methods concurrently handle nonstandard data structures, a mixture of data types, nonhomogeneous variable relationships, and different degrees of influence of the variables.

An overview of Breiman et al.'s methodology will be given in the context of allocating and distributing 155mm howitzer fire. Critiques of this proposed approach and suggestions for alternative approaches are invited.

- Fire Direction Officer [FDO] (determines or approves the number of rounds and the shell/fuze combination to fire on the target)
- Type/Subtype (description of the type of target)
  e.g., artillery/medium.
- Size (in meters),
- Method of Engagement (how to fire on the target)
  e.g., fire-for-effect when ready.
- Degree of Protection (position of the target)
  e.g., standing on first volley and laying down on subsequent volleys.
- Strength (number of units comprising the target)
- Target Speed (in kilometers per hour)
- Sensor (friendly unit sighting the target)
  e.g., forward observer.
- Sensor Speed (in kilometers per hour)
- Sensor to Target Range (in meters)
- 155mm Howitzer to Target Range (in meters)
- Ammunition Available (both as number of rounds available by munition type and as the initial ammunition load expressed as a percentage of a basic load)
  e.g., 100 rounds of high explosive rounds which is x% of a basic load.
- Allocation Method (method of firing the rounds on a target)
  e.g, fire high explosive and smoke rounds simultaneously on the target
  [as opposed to firing all high explosive rounds first followed
  by the smoke rounds].
- Total Number of Rounds Fired on the Target (number)
- Number of First Munition Rounds Fired
  e.g., 6 rounds of high explosive.
- Type of First Munition Fired
  e.g., high explosive.
- Number of Second Munition Rounds Fired
  e.g., 8 rounds of smoke.
- Type of Second Munition Fired
  e.g., smoke.

Figure 1. Information Available for Each Decision.

II. Background

a. Data Sets

In December 1985, the BRL conducted a controlled laboratory experiment, the Firepower Control Experiment [2], at the joint U.S. Human Engineering Laboratory and BRL Command Post Exercise Research Facility. As part of this statistically designed experiment, information was collected on Fire Direction Officers' (FDOs') decisions on a variety of targets being forwarded to 155mm howitzer units.* This data set comprises 3,219 FDOs' tactical fire control decisions collected for different FDOs, target types/subtypes, target sizes, types of fire mission control (i.e., "method of engagement") and initial ammunition basic loads.

As part of the BRL's research in tactical computer science, several unclassified scenarios between friendly and enemy forces in the Fulda Gap have been developed under a BRL contract with LB&M Associates, Inc., Lawton, OK. Embedded within these scenarios are decisions on allocating and distributing 155mm howitzer fire on independent targets observed in one-hour periods. To date, information associated with 522 tactical fire control decisions has been extracted from a portion of these scenarios.

Figure 1 summarizes the type of information available for the decisions in these data sets. A combination of categorical and numerical variables describes the principle factors thought to influence the decision process (FDO through ammunition available) as well as the actual decision (allocation method through type of second munition fired). Based on the results of previous data analyses, it is anticipated that these variables have different degrees of influence and exhibit nonhomogeneity.

b. Parametric and Nonparametric Procedures Applied

1. Multiple Regression Analysis

Multiple regression analysis [3] is an analytical methodology that usually has one of the following primary goals: 1) predict the value of the dependent variable for new values of the independent variables, 2) screen variables to detect each variable's degree of importance in explaining the variation in response, 3) specify the functional form of the model, or 4) provide estimates of each coefficient's magnitude and sign. By applying multiple regression analysis to the data from the Firepower Control Experiment, it was hoped that a regression equation could be derived to suitably predict the allocation method. Using a combination of indicator factors for the categorical variables (e.g., FDO and target type/subtype) and untransformed values for the numerical variables (e.g., ammunition load expressed as a percentage of a basic load, target size, and the method of engagement), stepwise and "best subset" regressions were run to predict the response factor (e.g., the allocation method).

*Tactical Fire Direction and gunnery instructors from the US Army Field Artillery School, Fort Sill, OK, participated as FDOs

Stepwise regression [4] was run to insert factors into the regression equation based on their partial correlation coefficient with the response factor. At each step, the partial F criterion of each regressor already in the equation was compared to the appropriate tabled F value. The regressor was either retained in the equation or rejected based on whether the test was significant or not. Stepping continued until none of the regressors could be removed, and none of the other potential regressors could be inserted due to the value of their partial correlation coefficient. "Best subset" regression was then run on the stepwise regressor variables to determine the best overall subset out of all possible regressions according to the maximum R^2 criterion.

As a consequence of performing a least squares fit of the data, fitted equations were obtained for the allocation method. However, based on the proportion of variance accounted for by the regressors in the regression equations, none of the factors clearly influenced the allocation method. This suggests that other factors not taken into account may influence FDOs' decisions on an allocation method.

2. Mann-Whitney test

One of the objectives of the experiment was to test whether the amount of available ammunition affected the number of rounds the FDO elected to fire on a target. Prior to comparing all FDOs within a given ammunition basic load or comparing an individual FDO across the three ammunition basic loads, it was desirable to first examine whether or not it would be necessary to distinguish between the adjust fire (AF) and fire-for-effect (FFE) methods of engagement. Since the distribution of total rounds fired against a target is not known for the two employed methods of engaging a target, the nonparametric Mann-Whitney test [5] was used to test whether the two independent random samples could have been drawn from two populations having similar distribution functions. Based on the results of the Mann-Whitney test, the samples associated with the two methods of engagement could not be grouped together for other statistical tests.

3. Kruskal-Wallis Test

Similar to the Mann-Whitney test, the nonparametric Kruskal-Wallis one-factor analysis of variance by ranks procedure [5] was used to examine, *first*, the mean number of rounds fired within each of the three different ammunition basic loads by each FDO, and, *second*, the mean number of rounds fired by each of the three FDOs within a given ammunition basic load. It was concluded from the test that there were significant differences within an ammunition basic load in the mean number of rounds fired by each FDO against an individual target. In addition, test results showed that only one of the FDOs tended to fire on average more rounds against a target under at least one of the ammunition basic loads than under at least one of the other basic loads. For the random samples resulting in rejection of the null hypotheses, i.e., no difference in the mean rounds fired against a single target, additional pairwise Kruskal-Wallis tests were performed.

4. Cluster Analysis

Cluster analysis [6] was employed to categorize targets according to their importance based on their contribution to an enemy force in a particular tactical situation, i.e., their target value [7]. There are several ways to measure the value of the target. For example, one way could be to use several variables to measure the description, location, and activity of the target. A description of the target might include its type/subtype, size, and degree of protection. The location of the target might consider the actual grid location of the target, the altitude of the target, and the distance between the target and specific friendly units. The activity of the target might take into account its velocity and direction of movement.

Cluster analysis provided a multivariate statistical method to examine the interrelationships between the target description, the FDOs, and the initial ammunition load expressed as a percentage of a basic load. Target value was based on the mean number of rounds expended against an individual target. Targets were categorized into three target value clusters, i.e., "low", "fair", or "high", based on the minimization of the Euclidean distance between each target and the mean of the targets in the cluster.

c. Deficiencies Among the Analyses

Despite the fact that each of these statistical procedures is well known and used, they have several shortcomings with regard to the problems inherent to the Firepower Control Experiment data set. For instance, these methods do not concurrently handle the nonstandard data structures, a mixture of data types, nonhomogeneous variable relationships, and different degrees of influence of the variables. Subsequently, it is expected some information has been lost.

Thus, the combined results of these procedures do not provide an effective means of allocating and distributing 155mm howitzer fire for enemy targets. For instance, cluster analysis provides a coarse evaluation of a target's value based on the initial ammunition load, its type/subtype, and FDO. The "best subset" multiple regression equations provide only weak relationships between the FDO, allocation method, target type, target size, method of engagement, and initial ammunition load. Thus, the question remains, "Is this a result of variables measured in the experiment or a consequence that these procedures could only be focused on limited subsets of the data collected?" Subsequently, a search for a different means of analyzing this data has been undertaken.

III. Classification Tree Methodology

a. Background

Trees, whether known as decision trees, binary trees, or by some other name, have been previously used by data analysts as an informative nonparametric tool for investigating various types of data sets. Tree classification methods use the data to form prediction rules for a response variable based on the values of independent variables. Specifically, measurements are made on some object, and a prediction rule is then used

to decide to what class the object belongs. This methodology is so simple that it is often passed over in favor of other methods which are thought to be more accurate, such as discriminant analysis.

Recent developments in the area of structured classification trees, which have been published by Breiman et al., are aimed at strengthening and extending the original tree methodology. Their advancements have been incorporated into a statistical software package known as CART™ (Classification and Regression Trees). Given complex data sets with many independent variables, the developers of CART feel that the structured trees produced by CART can have error rates that may be significantly lower than those produced by the usual parametric techniques. These procedures are robust, e.g., they minimize the effects that data outliers might produce.

We feel that the advancements made in the area of structured tree methodology are significant enough to warrant investigation and application to the problems of allocating and distributing 155mm howitzer fire.

b. Overview of the CART Methodology

1. Definitions

Many of the statistical techniques presently available are designed for small data sets having a standard data structure. By a standard data structure we mean that there are no missing values among the measurements made on an object, or so few they may be estimated prior to analyzing the data. In addition, the variables all have to be of the same type, i.e., all numerical or all categorical. The underlying assumption of the data is that the driving phenomenon is homogeneous, i.e., the *same relationship* holds over the entire set of measurements made on the object in question.

The data which is available to study the problem of allocating and distributing friendly fire on enemy targets does not meet the above criteria. In both data sets, values for several of the measurements used to describe an enemy target may be missing or must be assumed not available for any number of reasons. The variable list comprising the make-up of a target's description (to include such items as its location, activity, description, etc.) is a mixture of both numerical and categorical variable types. Finally, we cannot realistically expect the same relationships to hold amongst the wide range of measurements made on a target.

2. Constructing a Classification Tree

To initially construct a structured tree, four elements are needed: 1) a set of binary questions of the form: Is $\mathbf{x} \in A$?, $A \subset X$, where $\mathbf{x}$ is the measurement vector defining the measurements (z_1, z_2, ...) made on a case, and X is defined as the measurement space containing all possible measurements, 2) a goodness of split criterion that can numerically evaluate any split of any node of the tree, 3) a rule which dictates when to continue splitting the node or to declare it a terminal node, and 4) a rule for assigning

every terminal node to a class. The set of binary questions generates a set of splits of every node. Those cases answering "yes" go to a left descendant node, while those answering "no" go to a right descendant node.

3. Features and Advantages

Breiman et al.'s methodology for classification trees appears to be a powerful and flexible analytical tool. Some of its major features and advantages over other methods will be very briefly outlined.

One of the more important aspects of the CART methodology is its ability to automatically handle missing values while minimizing the loss of information. This is achieved via the concept of surrogate splitting.

To understand surrogate splitting, two splits are said to be associated at a node if either of two conditions exists. If most of the cases are sent to the left or to the right by one split, and the other split also sends most of the cases in the same direction, the two splits are said to be strongly associated. On the contrary, the splits are also associated when one split sends most of the cases to the left (right) while the other split sends most of the cases to the right (left). The missing value algorithm then proceeds as follows. The CART methodology is designed to initially search through all possible splits on a given node and select the best split. For example, suppose the best initial split is: Is $x(5) > 34.1$?. All other variables except $x(5)$ will then be searched until the split on each variable which is most closely associated with the split on $x(5)$ is found. This series of splits might result in a list such as the following

$x(2) > 26.2$ is the most closely associated with $x(5) > 34.1$

$x(11) > 50.6$ is the second most closely associated with $x(5) > 34.1$

and so forth. These splits are the surrogate splits for the initial split: Is $x(5) > 34.1$?.

If a case has a missing value of $x(5)$ so that the best split is not defined for that case, CART then looks at all nonmissing variables in that case and finds the one having the highest measure of predictive association with the best split. In this example, CART would first look at the most closely associated surrogate split. For example, if the value of $x(2)$ is not missing, then the case would go left if $x(2) > 26.2$ and right otherwise.

This procedure is analogous to the one used to estimate the missing values in a linear model (viz., regression on the nonmissing value most highly correlated with the missing value). However, the CART missing value algorithm is more robust. The cases with missing values in the selected splitting variable do not determine which direction the other cases will take. Since further splitting continues, there is always the possibility that cases which may have been sent in the wrong direction due to the missing value algorithm will still be classified correctly.

Since variables do not act alone when predicting a classification, it is natural to question which variables played the role of predictors. In the construction of a tree there may be instances in which some of the variables are never used to split any node; however, this does not necessarily mean these variables lack any predictive information. Therefore, each variable is assigned a measure of importance which may be helpful to the analyst in uncovering variables otherwise glossed over when looking at only the splits from the final selected tree. One note should be made. Like many variable ranking procedures, this one is a bit subjective and the exact numerical values should not be interpreted precisely.

Other features which do not require such an in-depth discussion are the following: 1) ability to handle both numerical and categorical variables in a natural and simple fashion, 2) application to any type of data structure through the formulation of an appropriate set of binary questions, 3) a variable selection process closely resembling a stepwise procedure since a search is made at each intermediate node for the most significant split, and 4) in the overall measurement space X, the trees exhibit a robustness property similar to medians, while within the learning set the method is not appreciably affected by several misclassified points.

c. Digit Recognition Example Using the CART Methodology

The following digit recognition example was constructed by the authors of CART and illustrates the various parts of the classification portion of the methodology.**

Most of us are familiar with electronic calculators which ordinarily represent the digits 1, ..., 9, and 0 using seven horizontal and vertical lights in specific on-off combinations. If the lights are numbered as shown in **Figure 2**, then i denotes the ith digit, $i = 1, 2, ..., 9$, and 0, and the measurement vector $(x_{i1}, ..., x_{i7})$ is a seven-dimensional vector of zeros and ones. Let $x_{im}=1$ if the light in the mth position is "on" for the ith digit, otherwise $x_{im}=0$. **Table 1** presents the possible values of x_{im}. Set the number of classes $C = \{1, ..., 10\}$ and let the measurement space X contain all possible 7-tuples of zeros and ones.

Suppose the data for this problem are generated from a faulty calculator for which it is known that each of the seven lights has the probability of 0.1 of not functioning properly. The data consist of outcomes from the random vector $(X_1, \ldots, X_7, Y)$ where Y is the class label and assumes the values 1, ..., 10 with equal probability and, as noted previously, the $X_1, ..., X_7$ are zero-one variables. Given Y, the $X_1, ..., X_7$ are independently equal to the value corresponding to Y in **Table 1** with probability of 0.9 and are in error with a probability of 0.1.

** It should be pointed out here that while this is the same example as outlined by the authors in their textbook, the output they produced for the purpose of illustration was not generated by the learning sample data presented in the text. Padraic Neville, who has been assisting the authors with the software management, has stated that the original data used to run this example was accidentally lost, however, the data in the text nearly depicts the original data. Therefore, the final structured tree presented in this paper will differ from that presented in the text.

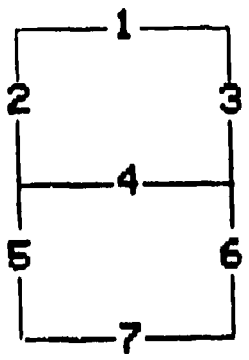


Figure 2. Horizontal and Vertical Lights.

Table 1. Possible Values of x_{im}.

Digit	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	0	0	1	0	0	1	0	1
2	1	0	1	1	1	0	1	2
3	1	0	1	1	0	1	1	3
4	0	1	1	1	0	1	0	4
5	1	1	0	1	0	1	1	5
6	1	1	0	1	1	1	1	6
7	1	0	1	0	0	1	0	7
8	1	1	1	1	1	1	1	8
9	1	1	1	1	0	1	1	9
0	1	1	1	0	1	1	1	10

The learning sample, L, is comprised of two hundred samples which are generated using the above distribution. Recall that each sample in L is of the general form $(x_1, ..., x_7, j)$ where $j \in C$ is the class label and the measurement vector $x_1, ..., x_7$ consists of zeros and ones.

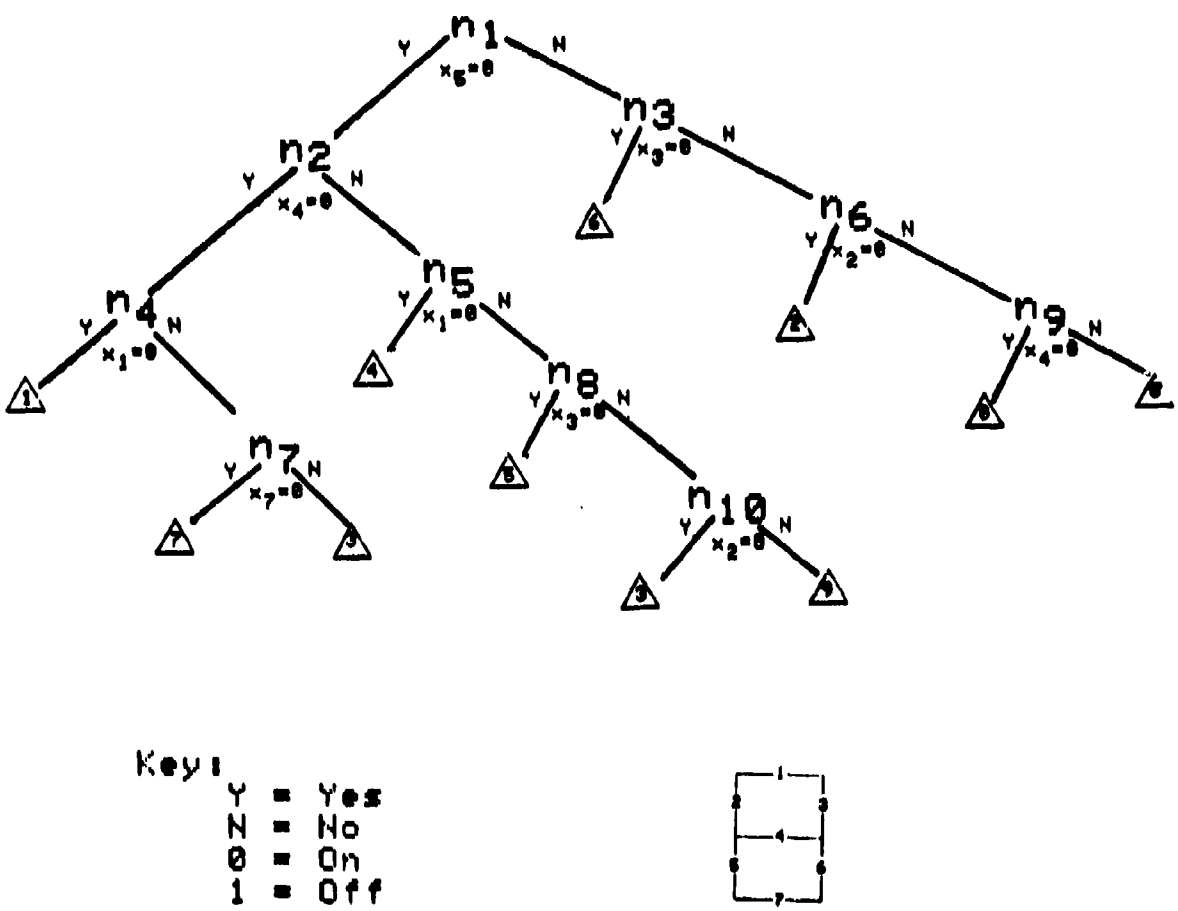
As previously mentioned in Section III.b.2., to apply the CART structured classification construction on L, four things must be specified: 1) the set of questions, 2) a rule for selecting the best split, 3) a criterion for choosing the right-sized tree, 4) a rule for assigning every terminal node to a class. Here the question set consisted of the seven questions: Is $x_m = 0$? where $m = 1, ..., 7$. The Gini index of diversity rule was used to select the best split. The concept of this splitting criterion depends on a node impurity measure. Given a node n with estimated class probabilities $p(j \mid n)$, $j = 1, ..., J$, and the probability that given a randomly selected case of unknown class falls into node n that it is classified as class i, define a measure $i(n)$ of the impurity of the given

node n as a nonnegative function ϕ of the $p(1 \mid n)$, ..., $p(J \mid n)$. Subsequently, the Gini index of diversity takes the form: $i(n) = \sum_{j \neq i} p(j \mid n)\, p(i \mid n)$. This node impurity is largest when all classes are equally mixed together in the node and smallest when the node contains only one class. A search is made for the split that most reduces the node, and consequently tree, impurity. The V-fold cross-validation method was used to "prune" to the right-sized tree. Here the original learning sample L was divided by random selection into V subsets L_v, $v = 1$, ..., V, of nearly equal size. The vth learning sample is: $L^{(v)} = L - L_v$, $v = 1$, ..., V, where $L^{(v)}$ contains the fraction $(V-1)/V$ of the total data cases (the cases in L but not in L_v). For example, if V is taken as 10, each learning sample $L^{(v)}$ contains 9/10 of the cases. Assume that a classifier can be constructed using any learning sample. Then, for every v, apply the classification procedure and let $d^{(v)}(\mathbf{x})$ be the resulting classifier. Since none of the cases in L_v was used to construct $d^{(v)}$ (the classifier), a sample estimate of the overall tree misclassification rate may be calculated, and a classifier is now constructed using the entire original learning sample L. The assignment rule proposed was to classify a terminal node n as that class for which $N_j(n)$ is largest, where $N_j(n)$ is the number of class j observations in n.

The resulting classification tree is shown in **Figure 3**.† The question leading to a split is indicated directly underneath each intermediate node. If the question is answered affirmatively, the split is to the left; if it is answered negatively, the split is to the right. Note that there are 11 terminal nodes, each corresponding to at least one class with class 3 having a second terminal node. Generally speaking, such a one-to-one correspondence occurs by accident since any number of terminal nodes may correspond to a particular class, or some classes may have no corresponding terminal nodes.

The overall probability of misclassifying a new sample given the constructed classifier (and the above fixed learning sample), $R^*(I)$, was estimated as 0.31. Two other estimates of $R^*(I)$ were also computed: 1) the cross-validation estimate, and 2) the resubstitution estimate. Since the learning sample, L, must be used in actual problems to construct both the classifier and to estimate $R^*(I)$, these estimates are referred to as internal estimates. The cross-validation estimate was estimated as 0.32 - satisfactorily close to $R^*(I)$. The resubstitution estimate was also calculated to be 0.32. This particular estimate identifies the proportion of cases from the learning sample, L, which is misclassified once the set is run through the constructed classifier. Using the V-fold cross-validation method explained earlier, such estimators come satisfactorily close to $R^*(I)$.

† The notation used here to describe the classification tree differs from that of the text



Key:

Y = Yes
N = No
0 = On
1 = Off

$n_1, \ldots, n_{10}$ = intermediate nodes

△ = terminal node

△₀, △₁, ... △₉ = terminal node classes

Figure 3. Digit Recognition Classification Tree.

IV. Summary

The classification tree structured methodology developed by Breiman et al. currently seems to be a viable approach to analyzing the available data sets. Although the regression tree portion of Breiman et al.'s methodology has not been examined in

detail, it also may be another means of analyzing this data. In the case of the data from the Firepower Control Experiment, it should be interesting to compare the results of the multiple regression analysis, Mann-Whitney test, Kruskal-Wallis tests, and cluster analysis to the CART results.

A critique of this proposed approach and suggestions for alternative approaches are invited.

References

[1] Breiman, L., Friedman, J., Olshen, R. and Stone, C., *"Classification and Regression Trees*, Belmont, California: Wadsworth, Inc., 1984.

[2] Winner, Wendy A., Brodeen, Ann E.M., and Smith, Jill H., *Test Design and Analysis: Firepower Control Experiment Part 12 of 12*, U.S. Army Ballistic Research Laboratory Memorandum Report, BRL-MR-3612, June 1987.

[3] Myers, Raymond H., *Classical and Modern Regression with Application*, Boston, Massachusetts: Duxbury Press, 1986.

[4] Draper, Norman R., and Smith, Harry, *Applied Regression Analysis*, New York, New York: John Wiley & Sons, Inc., 1981.

[5] Conover, W.J., *Practical Nonparametric Statistics*, New York, New York: John Wiley & Sons, Inc., 1976.

[6] Romesburg, H. Charles, *Cluster Analysis for Researchers*, London: Lifetime Learning Publications, 1984.

[7] FC6-20-2, "Targeting and Target Value Analysis," coordinating draft, Fort Sill, OK: US Army Field Artillery School, October 1984.

A SIMPLE MATHEMATICAL MODEL FOR THE SIMULATION OF IR BACKGROUNDS

Denis F. Strenzwilk, US Army Ballistic Research Laboratory

Michael P. Meredith, Biometrics Unit, Cornell University

Walter T. Federer, Mathematical Sciences Institute, Cornell University

ABSTRACT

At the US Army Ballistic Research Laboratory (BRL), Aberdeen Proving Ground, Md., weapon system analysts use background models in order to: 1) establish "clutter" thresholds for firing algorithms: and, 2) to study the masking and false alarm effect of background in their effort to evaluate the performance of various weapon systems. The BRL has received from US Army Engineer Waterways Experimental Station (WES) several large data bases comprised of blackbody temperatures derived from measurements obtained with an IR sensor. The sensor was mounted on a helicopter and scanned in the cross-track direction perpendicular to the direction of flight (in-track). The data consists of temperatures of scene elements (pixels) for a plowed field, a forested area, and a grassy field. The primary objective of this research is to provide a simple mathematical model which provides simulated data that are consistent with descriptive statistics from the original spatially correlated data base. Such statistics include the mean and standard deviation of temperature, and its 'energy spectrum'. The Mathematical Sciences Institute (MSI) at Cornell University have suggested time series models and a Spatial Moving Average (SMA) model as two approaches to the problem. One long term objective of this type of investigation is to construct a method for relating parameters in the model to physical constants. If successful, the model may then be extended over the diurnal cycle and seasons.

I. INTRODUCTION

BRL to date has modeled target signatures in a deterministic manner while background signatures have been treated stochastically. The deterministic model for target signatures is appropriate because under a particular set of conditions, the signature is rather well defined and is amenable to a single characterization. The case is not the same for backgrounds, which are many and varied. Thus, the general approach in modeling backgrounds has been to select a data set of a homogeneous scene, to extract pertinent statistics, such as, the mean temperature, the standard deviation, the 'energy spectrum', the correlation between pixels, etc., and finally, to develop a model, which can simulate a 'typical' background segment with these same statistics.

In most smart weapon simulations, the sensor scans across many square meters of background before any target is encountered. During this time, the sensor's signals are processed by a target discrimination circuit that usually includes some sort of adaptive threshold logic. Usually for this type of discrimination, the signal's Root-Mean-Square (RMS) average is developed as a measure of background 'clutter'. Target

detections occur when the instantaneous sensor output exceeds a threshold value that is proportional to the average of the output signal. The sensor's output signals produced by scanning the modeled background are thus used to provide a basis for setting the detection threshold; this is perhaps the most important function of the background. The stochastic background modeling approach currently being used at the BRL is based on a normal temperature assumption. It is quite well suited to provide a reasonable estimate of average clutter in many situations, even though the temperature distribution of the pixels is not normal. However, a background model also ought to include some provision for sources of false detection. The simple stochastic background model described here is clearly not capable of fulfilling this objective, for there is only a very remote possibility of a false alarm when the detection threshold is set to some multiple of the RMS signal. What is lacking is a means for incorporating some realistic scene features that would constitute possible sources for false alarms.

Given that a target signature model with a reasonable degree of fidelity is mated with a valid stochastic background signature model , it is possible to predict when and where a target detection is likely to occur. Probabilities of target detection can be inferred and the sensor/processor may be analyzed in terms of performance given a target encounter. This has been the BRL approach for many smart weapon simulations. A different approach must be taken if one wants to make some assessment of the smart weapon's capability for rejecting false targets. Ideally, the background infrared signature model used for this type of performance analysis ought to include a realistic characterization of individual scene elements that might confuse the target discrimination logic. Might it be possible to develop a background signature model that is predictive in nature and includes specific features that are potential false targets? BRL would like such a model if the development effort does not cost us too much, and more importantly if the proposed model does not require so many computer resources as to interfere with those needed for the performance simulation.

An alternative to "modeling" the background signatures either deterministically or stochastically would be to use actual scene measurements as inputs to the smart weapon sensor model. This would require that the measured background signatures be compatible with the sensor model in terms of viewing direction, detector wavelength band, and scene pixel size. Although the existing infrared background signature data base is rather extensive , very few of these sources have the requisite characteristics for smart weapons system evaluations that are currently being conducted. One source of data found to be generally compatible with the type of smart weapons that are being investigated at the BRL is the set of infrared scanner measurements of a rural area near Hunfeld, Germany made by the US Army Engineer Waterways Experiment Station (WES). For these measurements WES employed a helicopter-mounted Daedalus infrared scanner operating in the wavelength band of 8.5 to 12.5 micrometers. The scanner was flown over the test terrain at altitudes of 200 and 600 feet. The sizes of the corresponding ground resolution elements were roughly compatible with the 0.1 meter resolution that is optimum for the BRL's smart munition evaluation efforts, and the site of the measurements and the scene content is quite appropriate. The advantage of modeling this data set is that the model can be checked against the actual data in the simulation of a smart weapons concept.

Up to this point the discussion has been confined to simple scenes, e.g., a grassy field, a plowed field, a forested area, etc. Once a suitable model for a simple scene has

been developed, BRL wants to construct arbitrary scenes from these simple scenes. Thus a forested area of any desired size may be placed next to a plowed field. A road may be added to the scene. This compound scene with these three different kinds of textures could then be used in computer simulations of smart weapon concepts. All kinds of different compound scenes of arbitrary geometry and composition could be constructed from the models of the simple scenes. Thus the ability to construct compound scenes from simple scenes is a desideratum of the modeling effort.

II. DATA BASE

In this paper the time series models were applied to the data of the forested area. The data of the plowed field and grassy area have a similar format. The data base for the forested area is composed of 250 rows of temperatures. Each row contains 500 temperature pixels. Thus, for this data set there are 250 rows times 500 columns or 125,000 pixels of temperature. A row of data (500 pixels) represents one 'cross-track' scan of the sensor, which was mounted on a helicopter that flew in a direction perpendicular to the rows ('in-track') . After processing the data with ground truth information, it was concluded that at the 600 ft altitude the in-track (flight direction) dimension of the pixels was 0.3050m whereas the cross-track dimension was 0.1525m. The data are highly correlated both in-track and cross-track.

III. TIME SERIES MODEL

For each row of 500 observations a (p=1, q=1) autoregressive moving average model, ARMA(1,1) was fitted to the data. If the actual temperature observation was used to forecast the next pixel value for a complete row of simulated data, the forecasted data had the same spatial pattern and statistical characteristics as the actual data. If, however, the forecasted value was used to forecast the next pixel value in the row, the resulting set of forecasted values did not have the same pattern but did have the same characteristics. Thus, to preserve the spatial pattern in the time series approach, the actual data base would have to be used to make the forecasts. It was decided that for most applications it would suffice to have a model with the same statistical characteristics. Therefore, the actual observation of the temperature of the first pixel in each row was used to forecast the 2nd value and thereafter the forecasted value was used to forecast the next pixel value in the row. The ARMA used was

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} - \theta_1 a_{t-1} + a_t(\mu_a, \sigma_a), \qquad \text{III.1}$$

where

t equals 1,2,3,...,500

z_t temperature of t th pixel in row

$\tilde{z}_t$ temperature of t th pixel in row minus the mean, $(z_t - \mu)$

μ mean temperature of row

ϕ_1 autoregressive parameter of order one

θ_1 moving average parameter of order one

a_t random number for t th pixel from $N(\mu_a, \sigma_a^2)$, called residual or 'shock'

μ_a mean temperature of residuals

σ_e standard deviation of residuals

IV. ENERGY SPECTRUM

Let us represent the the two dimensional array of temperatures as a matrix, whose elements $T(l,m)$ are

$$T(l,m)=z_t^l, \qquad \text{IV.1}$$

where

z_t^l is the value of z_t in the lth row

m equals $0,1,2,\ldots,N_r - 1$

N_r is the number of pixels in a row (=500)

l equals $0,1,2,\ldots,N_c - 1$

N_c is the number of pixels in a column (=250).

t equals $m+1$

The discrete Fourier transform (DFT) for a row of temperatures is

$$Z^l(k)=\sum_{m=0}^{N_r-1} T(l,m)exp[-i(2\pi/N_r)mk], \qquad \text{IV.2}$$

where

k equals $0,1,2,\ldots,N_r-1$,

and for a column of temperatures is

$$Z_m(k)=\sum_{l=0}^{N_c-1} T(l,m)exp[-i(2\pi/N_c)lk], \qquad \text{IV.3}$$

where

k equals $0,1,2,\ldots,N_c-1$.

The frequency of a row f_r is

$$f_r=m/N_r\Delta_r \qquad \text{IV.4}$$

where

Δ_r is .1525m,

and the frequency of a column f_c is

$$f_c=l/N_c\Delta_c \qquad \text{IV.5}$$

where

Δ_c is .3050m.

The energy of the kth frequency in the lth row $S^l(k)$ is

$$S^r(k)=Z^r(k)Z^{r*}(k), \qquad \text{IV.6}$$

and the energy of the kth frequency in the *mth* row S_m is

$$S_m(k)=Z_m(k)Z_m^*(k), \qquad \text{IV.7}$$

where the symbol * denotes the complex conjugate. The cross-track energy spectrum and the in-track energy spectrum are a statistical measure of the correlation of the data, and result when $S^r(k)$ or $S_m(k)$ are plotted against frequency, respectively. (Zero frequency is excluded as the interest is in the the variation from the mean.)

The energy spectrum is symmetrical about the Nyquist frequency, which occurs at $f_r=.5/\Delta_r=3.279$ cycles per metre and at $f_c=.5/\Delta_c=1.639$ cycles per metre. Thus, it is common practice to multiply the energy of the *kth* frequency by a factor of two, and to plot the energy spectrum up to the Nyquist frequency. This convention was used in this paper.

In order to approximate an ensemble average by a spatial average, it is customary[1] to average $S^r(k)$ over the 250 rows and to average $S_m(k)$ over the 500 columns. Thus, the average energy of the *kth* frequency of the 250 rows $S^r(k)$ is

$$S^r=(1/250)\sum_{l=0}^{249} S^r(k), \qquad \text{IV.8}$$

and the average energy of the *kth* frequency of the 500 columns $S_c(k)$ is

$$S_c=(1/500)\sum_{m=0}^{499} S_m(k). \qquad \text{IV.9}$$

V. TWO DIMENSIONAL ARMA MODEL

The criterion for selecting a model was that its mean temperature, its standard deviation, and its energy spectrum, which measures the correlation in the temperature, be in good agreement with the data. The mean temperature and the standard deviation of the data were evaluated. The energy spectrum of the data was evaluated and plotted versus the frequency for the cross-track and in-track directions.

The first two dimensional (2D) model tried was to simulate the 250 rows of temperature by using Equation (III.1) and the appropriate parameter estimates for each row. The mean temperature and its standard deviation were in good agreement. The cross-track energy spectrum for the rows $S^r(k)$ was also in good agreement with the data since the ARMA model was fitted to the rows. However, the in-track energy spectrum for the columns $S_c(k)$ was not in agreement with the data. This was expected because nothing had been done to introduce correlation between adjacent rows. Several approaches based on using the temperatures in the row above to forecast the next forecast in the row below were suggested as a way of introducing correlation. None of these approaches was successful.

After inspection of the spatial temperature variation of several sets of adjacent rows, some trends were noticed. The first was that $T(l,m)$ and $T(l+1,m)$ had similar values and the second was that if $T(l,m+1)$ increased or decreased from $T(l,m)$, then

---

[1] La Rocca, Anthony J. and Witte, David J.,"Handbook of the Statistics of Various Terrain and Water (Ice) Backgrounds from Selected U.S. Locations(U)," DTIC Technical Report Number 139900-1-x, January 1980, pages 2-11 to 2-12.

$T(l+1,m+1)$ would show a similar increase or decrease from $T(l+1,m)$. Perhaps, the shock a_t^l that produced $T(l,m+1)$ was correlated with the shock a_t^{l+1} that produced $T(l+1,m+1)$. Based on this physical evidence, the assumption was made that a_t^l was related to a_t^{l+1} through a bivariate normal distribution $g(a_t^l, a_t^{l+1})$ given by

$$g(a_t^l, a_t^{l+1}) = \left(\frac{1}{2\pi\sigma_a^l \sigma_a^{l+1}\sqrt{1-\rho^2}} \right) \exp\left[-\frac{1}{2(1-\rho^2)} \left(\left(\frac{a_t^l}{\sigma_a^l} \right)^2 - 2\rho \frac{a_t^l}{\sigma_a^l} \frac{a_t^{l+1}}{\sigma_a^{l+1}} + \left(\frac{a_t^{l+1}}{\sigma_a^{l+1}} \right)^2 \right) \right], \quad \text{V.1}$$

where the means of the residuals μ_a^l do not appear since they are approximately equal to zero, and the correlation coefficient ρ has the range

$$-1 < \rho < +1. \quad \text{V.2}$$

The marginal probability density function (pdf) for a_t^l is

$$g_1(a_t^l) = N\left(0\ ,\ (\sigma_a^l)^2\right), \quad \text{V.3}$$

and the marginal pdf for a_t^{l+1} is

$$g_1(a_t^{l+1}) = N\left(0\ ,\ (\sigma_a^{l+1})^2\right). \quad \text{V.4}$$

The conditional distribution for a_t^{l+1} given a_t^l is

$$g_2(a_t^{l+1}/a_t^l) = N\left[\rho\left(\frac{\sigma_a^{l+1}}{\sigma_a^l} \right) a_t^l\ ,\ (\sigma_a^{l+1})^2(1-\rho^2) \right] \quad \text{V.5}$$

Now, the following procedure was used to find that value of ρ which minimized in the least squares sense the difference between the in-track energy spectrum of the data $S_c(k)$ and the in-track energy spectrum of the simulated data $S_c(k;\rho)$. For a given value of ρ the first row of simulated temperatures was generated from the ARMA model given in Equation (III.1) with the appropriate parameter estimates by using the values of a_t^0 drawn from the marginal distribution given in Equation (V.3). The second row of simulated temperatures was generated from the ARMA model given in Equation (III.1) with the appropriate parameter estimates by using the values of a_t^1 drawn from the conditional distribution given in Equation (V.5). The set of a_t^1's for the second row were then used to generate the a_t^2's for the third row through the conditional distribution given in Equation (V.5), etc., until 250 rows of simulated temperatures were generated. Then, the in-track energy spectrum $S_c(k;\rho)$ was evaluated. The process was repeated for several values of ρ and the sum of squares of differences between the in-track energy spectrum for the data and the simulated data was evaluated for each value of ρ. The value of ρ which minimized this sum was chosen as the ρ to be used in this model.

VI. CONCLUSIONS FOR 2D ARMA MODEL

The value of ρ which minimized the difference in the actual and simulated energy spectrum was 0.89. The mean temperature $\bar{T}$ of the data base was $13.1°C$ and its standard deviation σ was $1.2°C$, whereas the simulated data base had a mean temperature of $13.1°C$ and a standard deviation of $1.1°C$. The comparison of the cross-track energy spectrum for the data and for the simulated data can be seen in Figure 1. Similarly, the comparison of the in-track energy spectrum for the data and for the simulated data can be seen in Figure 2. The agreement in both cases is good. Thus, this two dimensional ARMA model can simulate the statistical characteristics of the data, but not the spatial variations. Furthermore, to obtain more than 250 rows use Row 249 parameter estimates for Row 251, Row 248 parameter estimates for Row 252, etc., and

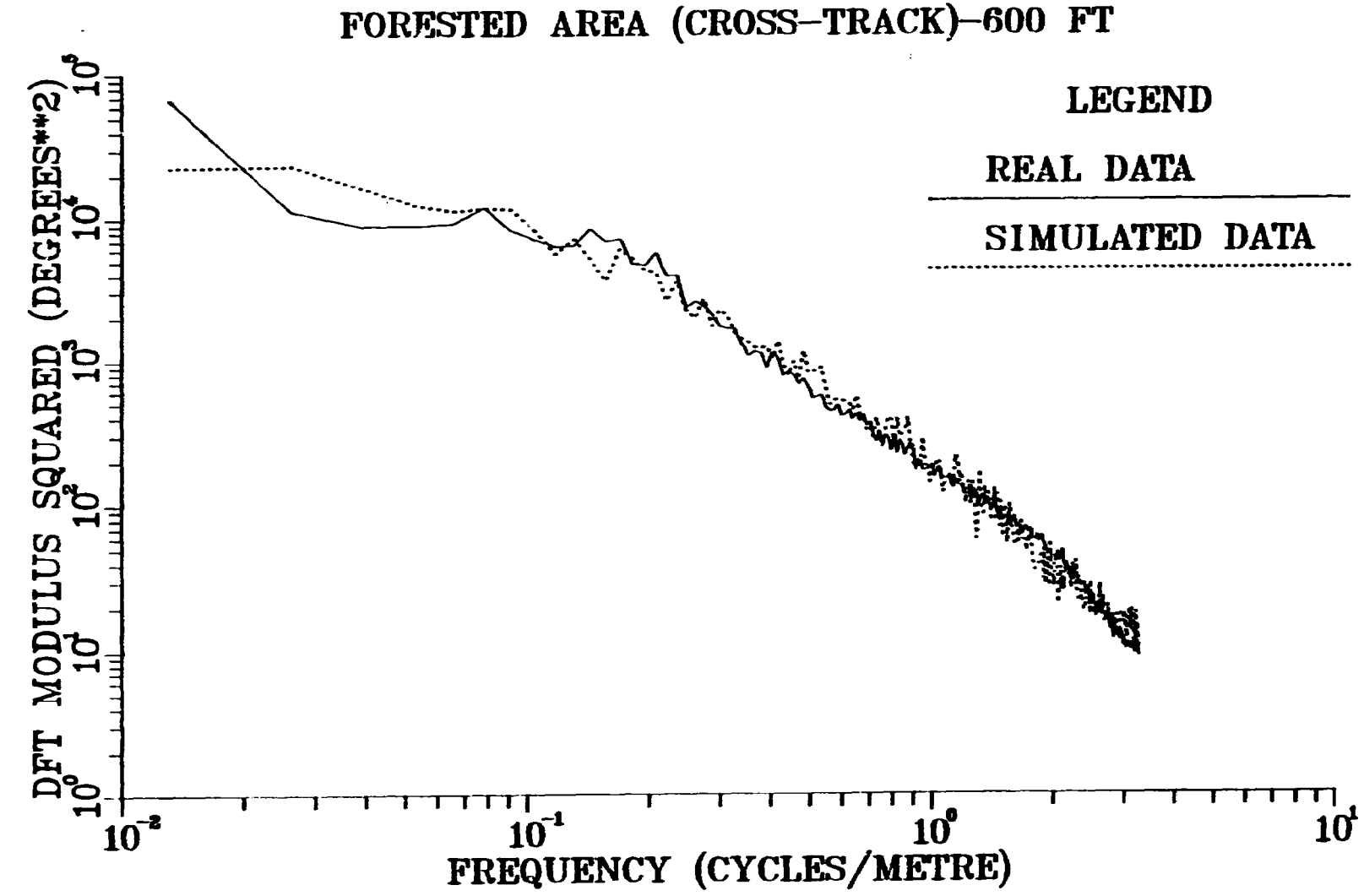


Figure 1. A comparison of the cross-track energy spectra of the real and simulated data of a forested area at 600 ft altitude is plotted versus frequency.

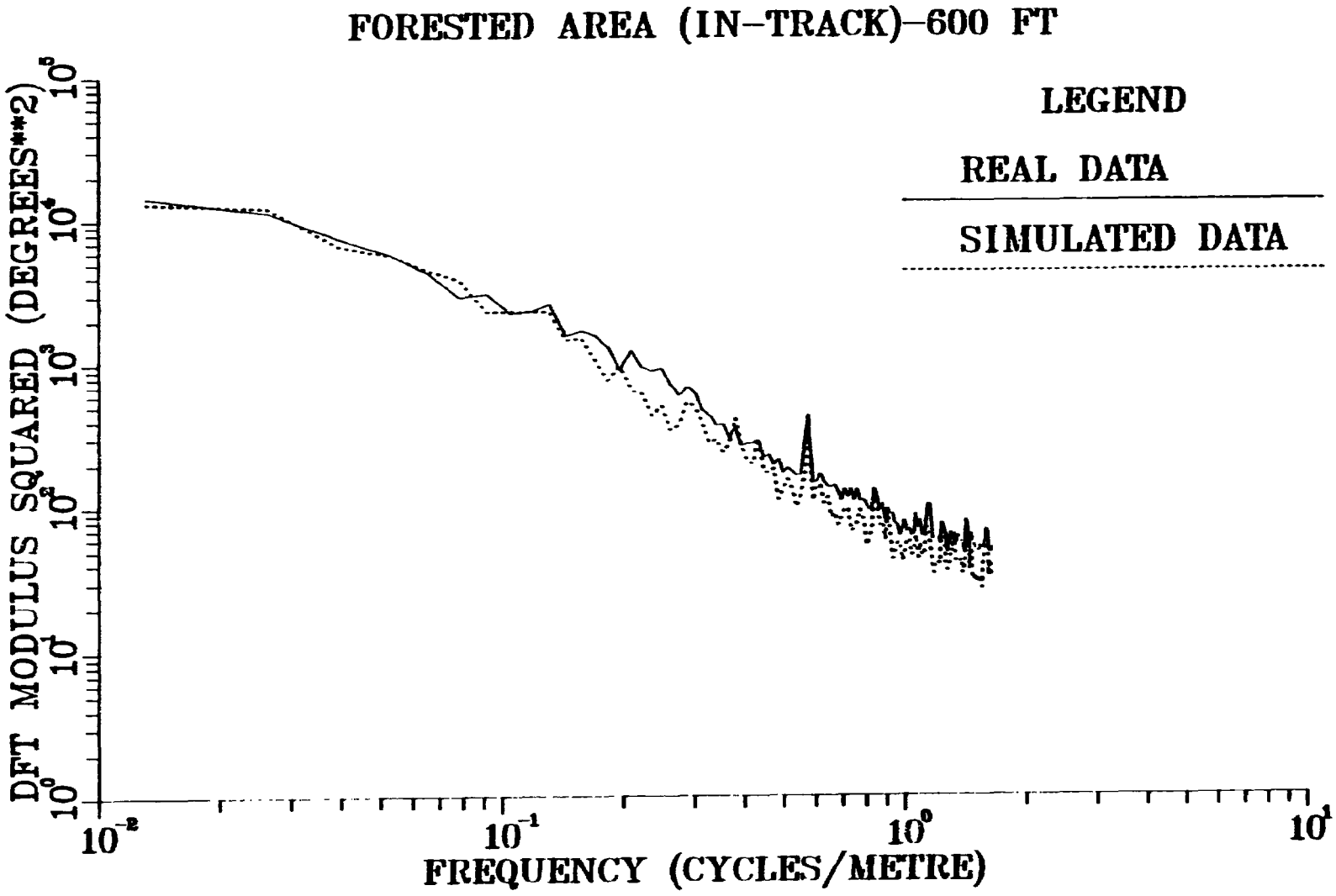


Figure 2. A comparison of the in-track energy spectra of the real and simulated data of a forested area at 600 ft altitude is plotted versus frequency.

essentially form a mirror image of the original 250 rows. To make rows longer, just draw more than 500 shocks for each row. An alternative to this procedure would be to use the 250 x 500 array of temperatures as the basic unit and extend it in any direction by mirror reflection.

One untried approach to improve this 2D ARMA model would be to take the average value of the ARMA parameter estimates for the 250 rows or at least several consecutive rows to obtain "representative parameter estimates". Then, randomly perturb these representative parameter estimates within their observed bounds for each row to be simulated, and proceed as before to determine a suitable value of ρ for the simulated temperatures.

Another untried approach to improve this 2D ARMA model might be to fit an ARMA model to every *kth* row of data. Use the appropriate parameter estimates for Rows $1, k+1, 2k+1, etc.$. For the rows in between 1 and k, use a weighted average for the parameter estimates, e.g., Row 2 values are $[(k-1)/k]$(value of Row 1) + $(1/k)$ (value of Row k), Row 3 values are $[(k-2)/k]$(value of Row 1) + $(2/k)$(value of Row k) , etc. (Note that a small amount of noise could be added to each value.) Proceed as before to determine a suitable value of ρ for the simulated temperatures.

VII. SPATIAL MOVING AVERAGE MODEL

The model described in this section differs from the ARMA models discussed above in that it is a two-dimensional model from the start whereas the others are one-dimensional models adjusted to give a two-dimensional array of spatially correlated observations. It also offers more promise of reproducing the spatial variation of the data, but at present it has not been applied to our problem. The steps for the SMA model are:

1. Generate an array of Z_{ij}, which are independent, identically

   distributed normal random variables, NIID$(0,\sigma^2)$.

2. Use Z_{ij} in a spatial moving average (SMA) to construct

   the temperature datum $T_{n,m}$ as

$$T_{n,m} = \overline{T} + \sum_{i=-p}^{p} \sum_{j=-q}^{q} A_{ij} Z_{n+i,m+j} \quad , \qquad \text{VII.1}$$

   where $E[T_{n,m}] = \overline{T}$,

   and

$$Cov(T_{n,m}, T_{n+s,m+t}) = 0, \quad \text{if} \quad |s| > p, \ |t| > q; \qquad \text{VII.2.a}$$

$$Cov(T_{n,m}, T_{n+s,m+t}) = \sigma^2 \sum_{i=-p}^{p} \sum_{j=-q}^{q} A_{ij}^2 , \quad \text{if } s=0, \ t=0; \qquad \text{VII.2.b}$$

   and

$$Cov(T_{n,m}, T_{n+s,m+t}) = \sigma^2 \sum_{i=-p+s}^{p} \sum_{j=-q+t}^{q} A_{ij} A_{i-s,j-t} \quad , \quad otherwise. \qquad \text{VII.2.c}$$

3. A_{ij} are chosen by the experimenter such that

$$\sum_i \sum_j A_{ij} = 1 \quad . \qquad \text{VII.3}$$

Table 1 illustrates the needed coefficients A_{ij} for $p=1, q=1$ that multiply the random variable Z_{nm} in order to obtain a value for $T_{n,m}$ in Equation (VII.1).

TABLE 1. Coefficients of the Spatial Moving Average for Constructing the Datum $T_{n,m}$ Using the NIID Random Variables Z_{ij}.

	m-1	m	m+1
n-1	$A_{-1,-1}$	$A_{-1,0}$	$A_{-1,1}$
n	$A_{0,-1}$	$A_{0,0}$	$A_{0,1}$
n+1	$A_{1,-1}$	$A_{1,0}$	$A_{1,1}$

Some A_{ij} may be chosen to be zero or some other value.

PROBLEM: Optimal determination of A_{ij} in SMA to match marginal spectra from observed process.

VIII. SOME COMMENTS

Our primary objective in this research was to provide a simple mathematical model which provides simulated data that are consistent with descriptive statistics from the original spatially correlated data base. Our 2D ARMA model met our criterion that its mean temperature, its standard deviation, and its energy spectrum, which measures the correlation in the temperature, be in good agreement with the data, even though it did not reproduce the spatial variation in the data. Our assumption that the shocks in adjacent rows be drawn from a bivariate normal distribution was the ingredient that introduced the necessary two dimensional spatial correlation in the simulated data. Some additional approaches for simplifying our 2D ARMA model, which were centered around reducing the number of ARMA parameter estimates needed for simulation, have been suggested in the text. In addition a spatial moving average model has been outlined as an alternative method for this problem.

Our 2D ARMA model is an improvement over the normal models that are currently being used at the BRL, especially since the time series approach naturally forecasts outlier temperatures (false alarms) that are found in the data. In time, after more data are analyzed by ARMA models, methods for relating the parameter estimates to physical constants will be found. If successful, the model may then be extended over the diurnal cycle and seasons. Also, for the theorists, an n-dimensional spatially correlated model is easily constructed.

EVALUATION OF CAMOUFLAGE PAINT GLOSS VERSUS DETECTION RANGE

George Anitole and Ronald L. Johnson
U. S. Army Belvoir Research, Development And Engineering Center
Fort Belvoir, Virginia 22060-5606

Christopher J. Neubert
U. S. Army Materiel Command
Alexandria, Virginia 22333-0001

ABSTRACT

To increase durability, the military has considered using a higher gloss camouflage paint. The field test and statistical analyses required to determine paint gloss effects upon range of detection are described. Five, 5/4-ton CUCV trucks were painted in the woodland U.S./German pattern with 1, 5, 10, 15, and 20 percent paint gloss. At least 30 observers per gloss level were individually driven towards two sites. The distance of correct detections were recorded. An analysis of variance with individual comparisons determined that detection range was significantly ($\alpha < 0.05$) greater, when higher gloss levels were compared with the standard one percent.

1.0 SECTION I - INTRODUCTION

The current camouflage paint specifications used by the U.S. Army call for a lusterless finish. This particular finish was originally selected for camouflage purposes because of its low visual reflectance characteristic. The lusterless finish is the result of a high pigment to binder ratio, and tends to mark and scuff easier than paint with a lower ratio and higher gloss finish. In addition, colors in a glossier finish appear more vivid than lusterless finishes which acquire a washed out appearance much sooner. These phenomena have been the object of concern from a camouflage standpoint, since the use of glossier paints would result in a longer lasting camouflage effect.[1/] However, the problem in using glossier paints is the potential of increased reflectance, hence detection. It was the purpose of this field test to determine statistically the effect increased paint gloss would have on the range of target detection in a woodland background.

2.0 SECTION II - EXPERIMENTAL DESIGN

2.1 Test Paint

Camouflage paints were purchased in five different degrees of specular gloss from the Enterprise Chemical Coatings Co. Wheeling, Illinois. The paints were produced in colors Green

383, Brown 383, and Black using paint specification MIL-E-52798A, in 1, 5, 10, 15, and 20% reflectance measured at 60° (1% is the current gloss of military paint). The gloss percentage spread was selected to provide a noticeable difference in reflection considering normal manufacturing tolerances. The 20% reflectance level was selected as the upper limit, since any greater reflectance was considered too shiny for military purposes. One gallon of each color, in each reflectance, was purchased for test and shipped to Ft. Devens, MA where the field evaluation took place.

2.2 Test Targets

Five, 5/4-ton, commercial utility combat vehicles (CUCVs) on loan from the Massachusetts National Guard were painted by Belvoir personnel at the Ft. Devens Maintenance Facility in the standard United States/German three color woodland pattern.

2.3 Test Sites

The study was conducted at the Turner Drop Zone, Ft. Devens, MA, a large cleared tract of land surrounded by a mix of coniferous and deciduous forest resembling a central European background. Two test vehicle location sites were selected. Site #1 was located on the western end of the drop zone, so that the morning sun shown directly upon the test vehicle. Site #2 was located on the eastern edge of the drop zone, so that the afternoon sun shown directly upon the test vehicle. An observation path, starting at the opposite end of the drop zone from the test vehicle location, was laid out for each site. These layouts followed zig-zag, random length directions toward the test sites, and afforded a continuous line-of-sight to their respective test vehicle locations. The paths were within a 30° to 40° cone from the targets, and were surveyed and marked at 50 meter intervals using random letter markers. The markers and distances from the test vehicle location sites are shown in Table 1.

Table 1

Distances of Markers to Test Vehicles on Sites #1 and #2

Site #1		Site #2	
ALPHABET MARKER	DISTANCE IN METERS ALONG PATH FROM STARTING POINT TO TARGET	ALPHABET MARKER	DISTANCE IN METERS ALONG PATH FROM STARTING POINT TO TARGET
C	1,173.70	L'	1,261.50
U	1,132.02	I '	1,230.74
A	1,088.51	D'	1,192.40
R	1,044.10	B'	1 153.65
G	1,015.03	W'	1,[illegible]6.90
O	989.27	T'	1,076.05
F	947.17	U	1,033.50
X	901.17	H	987.16
K	854.06	L	942.80
P	808.71	T	902.04
H	762.36	J	853.57
Z	723.52	R	811.07
Q	706.95	K	770.70
J	693.23	I	731.23
V	653.54	V	693.06
D	608.16	F	648.52
S	569.96	Z	602.61
N	536.46	E	561.59
T	497.44	N	517.36
W	457.13	X	473.04
M	416.47	D	426.61
L	376.99	Y	392.77
E	342.99	S	354.92
I	296.01	P	320.74
Y	260.15	M	297.81
B	219.07	A	277.02
L'	172.15	C	239.95
B'	126.89	O	202.56
P'	79.71	G	162.82
O'	27.65	B	125.71
		W	92.19
		Q	51.84

2.4 Test Subjects

A total of 153 enlisted soldiers from Ft. Devens served as ground observers. All personnel had at least 20/30 corrected vision and normal color vision. A minimum of 30 observers were used for each test vehicle, about evenly split per test site. Each observer was used only one time.

2.5 Data Generation

The test procedure for determining the detection distances of the five vehicles involved searching for the vehicles while traveling along the predetermined measured paths. Each ground observer started at the beginning of the observation path, i.e., marker C for site #1 and marker L for site #2. The observer rode in the back of an open 5/4- ton truck accompanied by a data collector. The truck traveled down the observation path at a very slow speed, about 3-5 mph. The observer was instructed to look for military targets in all directions except directly to his rear. When a possible target was detected, the observer informed the data collector and pointed to the target. The truck was immediately stopped, and the data collector sighted the pointed target. If the sighting was correct i.e., the painted CUCV, the data collector recorded the alphabetical marker nearest the truck. If the detection was not correct, the data collector informed the observer to continue looking, and the truck proceeded down the observation path. This search process was repeated until the correct target was located.

The target CUCVs were rotated between the two test sites on a daily basis, until all vehicles had been observed by at least 15 observers at each site. Their orientations with respect to the sun were kept constant at both test sites. The vehicle side windows were left open to eliminate shine, and a tarpaulin was used to cover the windshield and rear window. The vehicles were positioned so that the left side was facing the direction of observer approach.

3.0 SECTION III-RESULTS

Tables 2, 3, and 4 show the detection data for the 5/4-ton CUCVs painted in 1, 5, 10, 15, and 20% gloss. Table 2 gives the mean detection range in meters for each gloss level, and its associated 95% confidence interval. Table 3 shows the analysis of variance[2/] performed upon the data of Table 2 to determine if there were significant differences in the detection ranges i.e., gloss has an effect upon detection range. Table 4 indicates which gloss levels differed significantly from each other. Figure 1 is a graphic display of the detection ranges of Table 2.

Table 2

Mean Gloss Detection Ranges (Meters) and 95 Percent Confidence Intervals.

% GLOSS LEVEL	N	MEAN	STANDARD ERROR	95 PERCENT CONFIDENCE INTERVAL LOWER LIMIT	UPPER LIMIT
1	31	580.0000	138.3944	529.2433	630.7567
5	30	790.1333	216.3083	709.3715	870.8951
10	31	971.0000	117.7326	927.0429	1014.9571
15	30	1078.3333	114.1195	1035.7252	1120.9415
20	31	1153.9677	93.1967	1119.7875	1188.1480

Table 3

Analysis of Variance for Vehicle Detection Across Five Levels of Paint Gloss

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	F-TEST	SIG LEVEL
GLOSS	4	6,611,277.3660	1652819.3415	81.7597	0.00000*
ERROR	148	2,971,691.1011	20215.5857		
TOTAL	152	9,582,968.4671			

BARTLETT'S TEST FOR HOMOGENEOUS VARIANCES

NUMBER DEGREES OF FREEDOM = 4,
F = 6.49661911766 SIGNIFICANCE LEVEL α = 0.0003
*Significant at α less than 0.001 level.

Table 3 indicates that there are significant differences in the ability of the ground observers to detect 5/4-ton CUCVs of different degrees of paint gloss. The Bartlett's Test indicates that the variances for each level of paint gloss are not homogeneous, i.e., significantly different, so they are not necessarily from the same population.

Table 4

Individual Comparisons Identifying Which Levels of Paint Gloss Differed Significantly from Each Other

1% Gloss and 5% Gloss
COMPARISON = -210.13333 SUM OF SQUARES = 673198.30383
F = 33.301 SIGNIFICANCE LEVEL = 0.00000 ***
1% Gloss and 10% Gloss
COMPARISON = -391.00000 SUM OF SQUARES = 2330808.68852
F = 115.298 SIGNIFICANCE LEVEL = 0.00000 ***
1% Gloss and 15% Gloss
COMPARISON = -498.33333 SUM OF SQUARES = 3786107.92350
F = 187.287 SIGNIFICANCE LEVEL = 0.00000 ***
1% Gloss and 20% Gloss
COMPARISON = -573.96774 SUM OF SQUARES = 5106304.01613
F = 252.592 SIGNIFICANCE LEVEL = 0.00000 ***
5% Gloss and 10% Gloss
COMPARISON = -180.86667 SUM OF SQUARES = 490691.26667
F = 24.273 SIGNIFICANCE LEVEL = 0.00000 ***
5% Gloss and 15% Gloss
COMPARISON = -288.20000 SUM OF SQUARES = 1245888.60000
F = 61.630 SIGNIFICANCE LEVEL = 0.00000 ***
5% Gloss and 20% Gloss
COMPARISON = -363.83441 SUM OF SQUARES = 2018183.50002
F = 99.833 SIGNIFICANCE LEVEL = 0.00000 ***
10% Gloss and 15% Gloss
COMPARISON = -107.33333 SUM OF SQUARES = 172806.66667
F = 8.548 SIGNIFICANCE LEVEL = 0.00348 **
10% Gloss and 20% Gloss
COMPARISON = -182.96774 SUM OF SQUARES = 510390.01586
F = 25.247 SIGNIFICANCE LEVEL = 0.00000 ***
15% Gloss and 20% Gloss
COMPARISON = -75.63441 SUM OF SQUARES = 87215.15248
F = 4.314 SIGNIFICANCE LEVEL = 0.03779 *

The following levels of paint gloss differed significantly from each other: 1% vs. 5%, 1% vs. 10%, 1% vs. 15%, 1% vs. 20%, 5% vs. 10%, 5% vs. 15%, 5% vs. 20%, 10% vs. 15%, 10% vs. 20% and 15% vs. 20%.

* Significant at α less than 0.05 level

** Significant at α less than 0.01 level

*** Significant at α less than 0.001 level

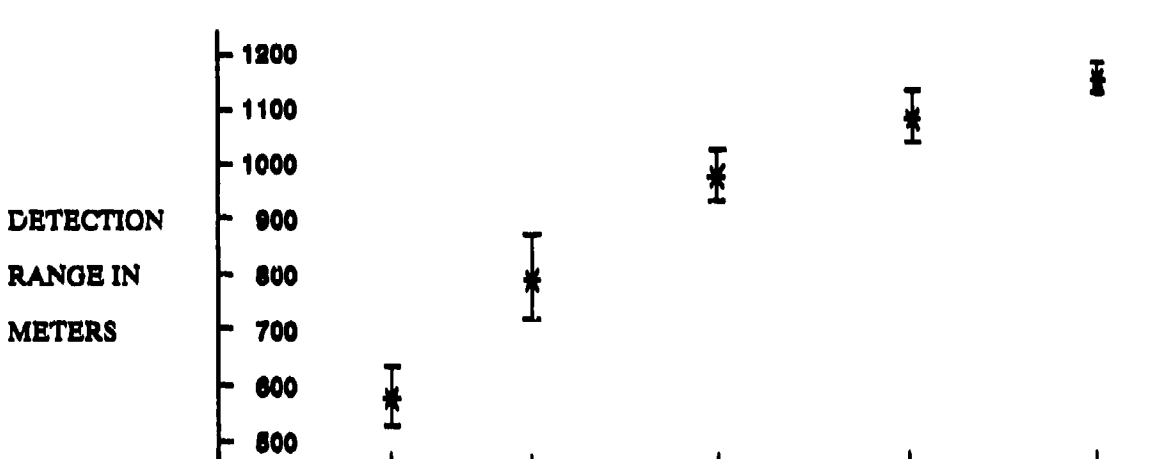


Figure 1. Detection Range In Meters for CUCVs Painted in Five Levels of Gloss

The Bartlett's Test for homogeneity of variance was significant at less than $\alpha = 0.001$. Thus, it can not be assumed that all the sample variances are from the same population. This assumption is required to perform the parametric test of analysis of variance and associated individual comparisons. When the Bartlett's Test is significant, non-parametric tests should be used to determine the relative positioning of the sample statistics. Two such non-parametric tests were performed, the Krushkal-Wallis One-Way Analysis of Variance and the Mann-Whitney U Test[3/]. The Krushkal-Wallis Test determined that there were significant differences between the levels of paint gloss. The Mann-Whitney U Test, based upon the Chi-Square distribution, determined the probability of individual gloss percentages differing from each other. These tests, while not as powerful as the parametric test, yielded the same general results, and are available upon request from the U.S. Army Belvoir Research, Development and Engineering Center, ATTN: STRBE-JDS, Fort Belvoir, VA 22060. It is not unexpected that the variances for each gloss level were not homogeneous. Each level of gloss was different from the preceding by 5%. These equal differences in shine are not perceived as such by the human eye. The 1% gloss was seen as dull, however the 5 through 20% paint gloss was perceived as being reflective. This is verified by viewing the differences in mean detection for the gloss percentages of 1 vs. 5, 5 vs. 10, 10 vs. 15, and 15 vs. 20 (see Table 5). If the variances were normally distributed, the mean differences between percentages of gloss would be about the same.

Table 5

Mean Differences In Detection Range (Meters) Between Gloss Levels

% GLOSS	MEAN DETECTION RANGE		DIFFERENCE
1 vs. 5	580	790	210
5 vs. 10	790	971	181
10 vs. 15	971	1078	107
15 vs. 20	1078	1153	75

4.0 SECTION IV - DISCUSSION

Figure 1 and Tables 2 through 4 clearly show that the higher the percentages of paint gloss, the longer the mean range of target detection. The differences between the 1% gloss detection range, and the 5, 10, 15, and 20% gloss detection ranges are significant well beyond the $\alpha = 0.05$ level. This α value is the probability that one will make a decision that the levels of paint gloss are significantly different in the resulting detection ranges when they are not. For this study, the decision is that the higher gloss paint levels of 5, 10, 15, and 20% will have a longer range of target detection than the 1% paint gloss level. In the world of statistics, if a decision has a probability of being wrong 5 or less times out of 100 ($\alpha = 0.05$) then this is an acceptable risk. If this probability of being wrong is greater than 5 times out of 100, the risk is not acceptable, and the decision is rejected. In the present study, these levels of differences in mean detection ranges tend to get smaller as the percentage of paint gloss increases (Figure 1 and Tables 2 and 4), but they never exceed the $\alpha = 0.05$ level. With the exception of the paint gloss comparisons 10 vs. 15% and 15 vs. 20%, which are significant at $\alpha = 0.003$ and 0.037 respectively, the other comparisons are significant at an α level less than 0.001. The differences between the detection means asymptotes as the percentage of the gloss gets higher (see Figure 1). This is due to the fact that targets with a higher gloss are easier to see than targets with a lower gloss. For example, increasing the paint gloss from 1 to 5% would increase the mean detection range by 210 meters (Table 5).

It was also observed that as the level of paint gloss increased, the visual perception of a pattern decreased. The camouflage pattern was difficult to discern at paint gloss levels of 10% and above.

5.0 SECTION V- SUMMARY AND CONCLUSIONS

Five 5/4-ton CUCVs were painted in the standard woodland United States/German three color pattern with the following paint glosses:

- 1% (standard)
- 5%
- 10%
- 15%
- 20%

A minimum of 30 ground observers per paint gloss level were driven toward each of two sites on marked observation trails in the back of an open 5/4-ton truck. The subjects were looking for military targets, and they informed the data collector when they thought they saw one. If the detection was correct, the closest alphabetic ground marker to the truck was recorded. From this letter, the exact distance to the target from the truck was determined. If the detection was not correct, the search continued with the truck traveling down the observation path until the test target was seen. An analysis of the resulting data provided the following conclusions:

A. The targets with the higher paint gloss of 5, 10, 15, and 20% were significantly easier to detect than the target with the 1% paint gloss.*

B. The higher gloss paint levels of 5, 10, 15, and 20% will have a significantly longer range of target detection than will the 1% paint gloss level, which will increase their vulnerability to enemy fire.

C. In that the 5% paint gloss vehicle was detected, on the average, 210 meters farther away than the 1% paint gloss vehicle, one can not recommend any increase in the paint gloss over the 1% currently being employed by the U.S. military.

* Low visual reflectance is particularly important in woodland backgrounds where reflection and brightness are relatively low. Its effect in bright backgrounds such as desert or arctic environments, where reflections from glossier paints may be lost in the noise, remains to be evaluated.

REFERENCES

1. Anitole, George and Johnson, Ronald, Saudi Arabian National Guard Camouflage Paint Development Program, U.S. Army Mobility Equipment Research and Development Command, Fort Belvoir, VA, 15 October 82.

2. Natrella, Mary G., Experimental Statistics, National Bureau of Standards Handbook 91, U.S. Department of Commerce, Washington, D.C., 1966.

3. Siegel, Sidney, Non-parametric Statistics For the Behavioral Sciences, McGraw-Hill Book Company, Inc., 1956.

SENSITIVITY ANALYSIS OF A NONSTOCHASTIC MODEL

A.A. Khan
US Army Concepts Analysis Agency
Bethesda, MD 20814-2797

ABSTRACT. Simulation models are now widely used as analytical tools. New models are usually subjected to quality assurance criteria before they can be employed in studies. This practice is prudent as well as useful in learning the characteristics of a newly developed simulation model. Also, it is necessary to find those parameters which have a significant impact on the response variable [1].

Mobilization Based Requirements Model (MOBREM), the model examined in this article will be used for policy studies and budget planning. Before it can be so employed , we subjected it to sensitivity analysis. Since the model is deterministic, there are no random errors in the response variable; therefore, the usual statistical methods are not applicable. In their place, the 'summary statistics' R^2 has been used judgmentally.

1-0 INTRODUCTION. The results in this report deal with the sensitivity analysis of the simulation model, Mobilization Based Requirements Model (MOBREM). This model has been designed to provide the U.S. Army with 'a responsive, consistent, and auditable system for determining the Continental United States (CONUS) resources required to support mobilization' [2]. This model was developed over a five year, five-phased period, from 1979 to 1984. It was delivered to Concepts Analysis Agency (CAA) in August 1984. Since then, the model has been used for the training of operators and for performing policy studies in connection with mobilization.

1-1 Sensitivity Analysis. A new model, before it can be used for any study, must be tested for its sensitivity to input parameters. In this report, we address the following issues:

a. From a selected list of input parameters (or factors), find those parameters which have a significant impact on the response variable.

b. Rank order the significant input parameters.

The response variable in this study is the manpower requirements by the major Army Commands (MACOMS) Installations, by Army Functional Dictionary (AFD) code, and by time periods from Mobilization day (M-day) to day of hostilities (D-day).

1-2 Background. MOBREM is a very large and complex simulation model. For our purpose it is essential to keep in mind that it is a deterministic model. There are no random number generators in the subroutines or modules. Repeated observations do not provide estimate of 'variance'. If we repeat an experiment with fixed input values, we do not get a new value for a response variable. For this reason the classical statistical procedures have to be modified to meet the specific situation of MOBREM. In particular, F-test and t-test are not valid. We use R^2, the coefficient of determination, as the index of goodness of procedures used in our analysis..

2-0 OVERVIEW OF MOBREM. It will help in understanding the objectives of this study to have some perspective in mobilizing large numbers of people. To provide the reader with the magnitude of the numbers involved, we present in Table 1 the initial and final stages of mobilization in MOBREM. We will skip the details of organizational complexities and the organizations which are required to manage this operation.

2-1 CONUS Base. The major functions of CONUS Base organizations are to provide the support that enable units to be deployed, trainees to be be trained, and

equipment and supplies to be shipped to the theater or within CONUS. They also provide medical support for theater medical evacuees as well as those patient loads generated in CONUS installations [2].

2-2 Projections. A profile of organizations in CONUS in peace and war is given below. It illustrates the staggering magnitude of manpower involved from the initial to the final phase of mobilization. The organizational complexities to synchronize various phases of this process quantitatively is the most important function of MOBREM, but will not be discussed here.

Table 1
CONUS Base Organizations

Units	Peacetime Strengths (000)	Wartime Strengths (000)
TDA		
OSA and OCSA	3.7	6.8
Joint and DEF ACTV	6.7	7.1
OSA and ARSTAF FOA	46.7	46.0
Commands in CONUS	347.6	658.7
Army Reserves	25.8	0
National Guard	20.4	0
TOE		
Training division	32.0	52.9
Training spt units	4.1	4.5
GSF units	29.8	37.1
Sep inf bde	19.0	20.1
Other	3.9	4.1
Totals	539.7	837.3

Table(s) of allowances (TDA) is the number of slots allocated to different organizations, it includes both civilian and military, and table(s) of organization and equipment (TOE), i.e., the number of personnel authorized to keep a unit of army functional.

3-0 DESIGN OF EXPERIMENT. The initial list of 30 parameters was pared down to 9 for this study to economize on computer time; since each run of MOBREM takes about 12 hours to complete. The selection of the final list of input parameters and their levels was carried out with the help of both civilian and military analysts.

3-1. Choice of Design: A two-level fractional factorial design was planned for sensitivity analysis. The full design was completed in two stages. In the first stage, the 9 factors included both scalar and matrix inputs. The non-scalar inputs were treated as scalars by the following convention:

High value	+ C.V
Low value	-C.V

where C is a constant, V is a non-scalar. In this way the design is the usual fractional factorial design. At the initial stage of the study, we are interested only in 'sensitive' parameters, their interactions are of less importance. By 'sensitive,' we mean those inputs which produce a large impact on the response variable. A Plackett-Burman (P-B) design was deemed most suitable in this phase [4]. The 9 parameters are listed below:

FACTOR	DESCRIPTION
A	M-Day to D-Day
B	Work week
C	Training load
D	Show rates
E	Hospital rates
F	Deploying MTOE levels
G	Non-deploying MTOE levels
H	TDA levels
I	Other levels

Only Factors A and D are scalars

The smallest P-B design to accommodate 9 parameters is a 12 run design given below . A P-B design allows us to assess the impact of the main effects, which in this layout are not confounded with higher order interactions [5].

Table 2
PLACKETT-BURMAN DESIGN
I STAGE
PACKAGES

RUN	A	B	C	D	E	F	G	H	I
1	+	-	+	-	-	-	+	+	+
2	+	+	-	+	-	-	-	+	+
3	-	+	+	-	+	-	-	-	+
4	+	-	+	+	-	+	-	-	-
5	+	+	-	+	+	-	+	-	-
6	+	+	+	-	+	+	-	+	-
7	-	+	+	+	-	+	+	-	+
8	-	-	+	+	+	-	+	+	-
9	-	-	-	+	+	+	-	+	+
10	+	-	-	-	+	+	+	-	+
11	-	+	-	-	-	+	+	+	-
12	-	-	-	-	-	-	-	-	-

+ HIGH LEVEL
- LOW LEVEL

'PACKAGE' stands for a policy, i.e., a particular combination of input values.

3-2. Second Stage Design. At the first stage, results showed that only 5 factors were important enough for further investigation. These are:

Table 3

FACTOR	DESCRIPTION
A	D-Day to D-Day
C1	Training load
C2	Training equipment
H1	TDA fill
H2	TDA equipment

H2 is the corresponding level of equipment allowed to the unit. In this scheme, all parameters are scalars and the second stage P-B design is shown in Table 4.

Table 4
P-B DESIGN
II STAGE

Run	A	C1	C2	H1	H2
13	-	-	-	-	-
14	+	+	+	+	-
15	+	+	+	-	+
16	+	+	-	+	+
17	+	-	+	+	+
18	-	+	+	+	+
19	+	+	-	-	-
20	+	-	+	-	-
21	+	-	-	-	+
22	+	-	-	+	-
23	-	+	+	-	-
24	-	+	-	+	-
25	-	+	-	-	+
26	-	-	+	+	-
27	-	-	+	-	+
28	-	-	-	+	+

+ HIGH LEVEL
- LOW LEVEL

4-0 LINEAR MODEL. The collection and analysis of data depends on the mathematical model which we postulate to explain the relationship between the response and the input factors. The selection of a fractional factorial design at two levels, a resolution III design (P-B Design), was made with the object of estimating the main effects; higher order interactions can be sacrificed at this stage. The reasons can be summarized as follows [6]:

- Not much is known about the model on how different inputs impact on the output.
- In this situation it is best to assume a linear model.
- All experiments under uncertain conditions are conducted with some risk. If later, it is found that interactions are more important, one can re-run the simulation model to obtain additional observations. Simulation models can be run anytime one chooses to do so, provided time and resources are not prohibitive.
- Simpler mathematical models help in clearer exposition of the conclusions.

4-1 Analysis. At this stage the assumptions of linearity and additivity are convenient to model our results. If the experimental region is not large, higher order interactions need not be included in the expression connecting the response to the input [7]. We approximate the functional relationship between the response y and the input factors x1, x2, ..., x9 by Taylor's expansion.

$$y = A_0 + A_1 x_1 + A_2 x_2 + .. + A_9 x_9 + R \qquad (1)$$

where A_i ($i = 0, 1, 2, ..., 9$) are unknown constants and R is the remainder term in the Taylor's series expansion.. Observe that this model does not have stochastic components and therefore statistical techniques cannot be applied. We use the least square (l.s.) methods in the estimation of A_i and use R^2 to measure the adequacy of the model (1). For a clear discussion of two-level fractional design and the techniques of estimation of main effects, we refer to [8]. The least square technique is used in (1) to evaluate and partition the total sum of squares into the component sum of squares. Each component is attributable to a specific factor, plus the sum of square due to the remainder term. This analysis is carried out for the data in the first stage. A typical run with the response variable at each time period is shown in Table 5.

Table 5

Time from M-Day	Total Manpower Requirements
M + 10	318671
M + 20	314747
M + 30	354932
M + 40	367936
M + 50	403887
M + 60	442291
M + 90	479470
M + 120	498009
M + 150	504354
M + 180	501839
M + 210	497962
M + 240	497845
M + 270	497494
MOB-AV	532915

Since there is an ANOVA at each time period and for each run, there are 13 x 12 = 156 ANOVAs. These are not listed here, but the result of the analysis is shown in Table 6, showing the ranks of the factors in descending order.

Table 6

Ranking of Packages in Descending Order

Factor	Package
B	Workweek
C	Training
H	TDA
A	M-Day to D-Day
G	Non-deploying MTOE levels
F	Deploying MTOE levels
D	Show rate
E	Hospital
I	Other Personnel

Visual analysis at this stage is most effective, Figure 1 shows the response variable against time, when grouped according to the levels of Factor B (workweek). Factor B is the driver of the manpower requirements, a result confirmed by the usual ANOVA techniques. Figure 2 clearly indicates the main effects which have clear impact on the response variable. Apart from B, A and C produce measurable impact on manpower requirements up to time M + 100, after that the effects of these factors is dampened out. Other factors have negligible effects as can be seen by inspecting Figure 3. This combination of ANOVA, graphs of main effects and aggregating results by each level of Factor B is carried out for a selected group of AFD's. The results confirm the hypothesis that the ranking in Table 6 is valid for the sampled AFD aggregations. This simple computer intensive graphical technique has been extensively used in this study.

4-2 II Stage Analysis. Since the workweek parameter is so decisive, no further investigation is required to measure the sensitivity of the response variable to this parameter at this stage. In the II stage of design, a 60-hour workweek was fixed. The number of input factors was narrowed to 5 factors. Again, a resolution III design was used to generate simulation data. The factors in the II stage design are given in Table 7.

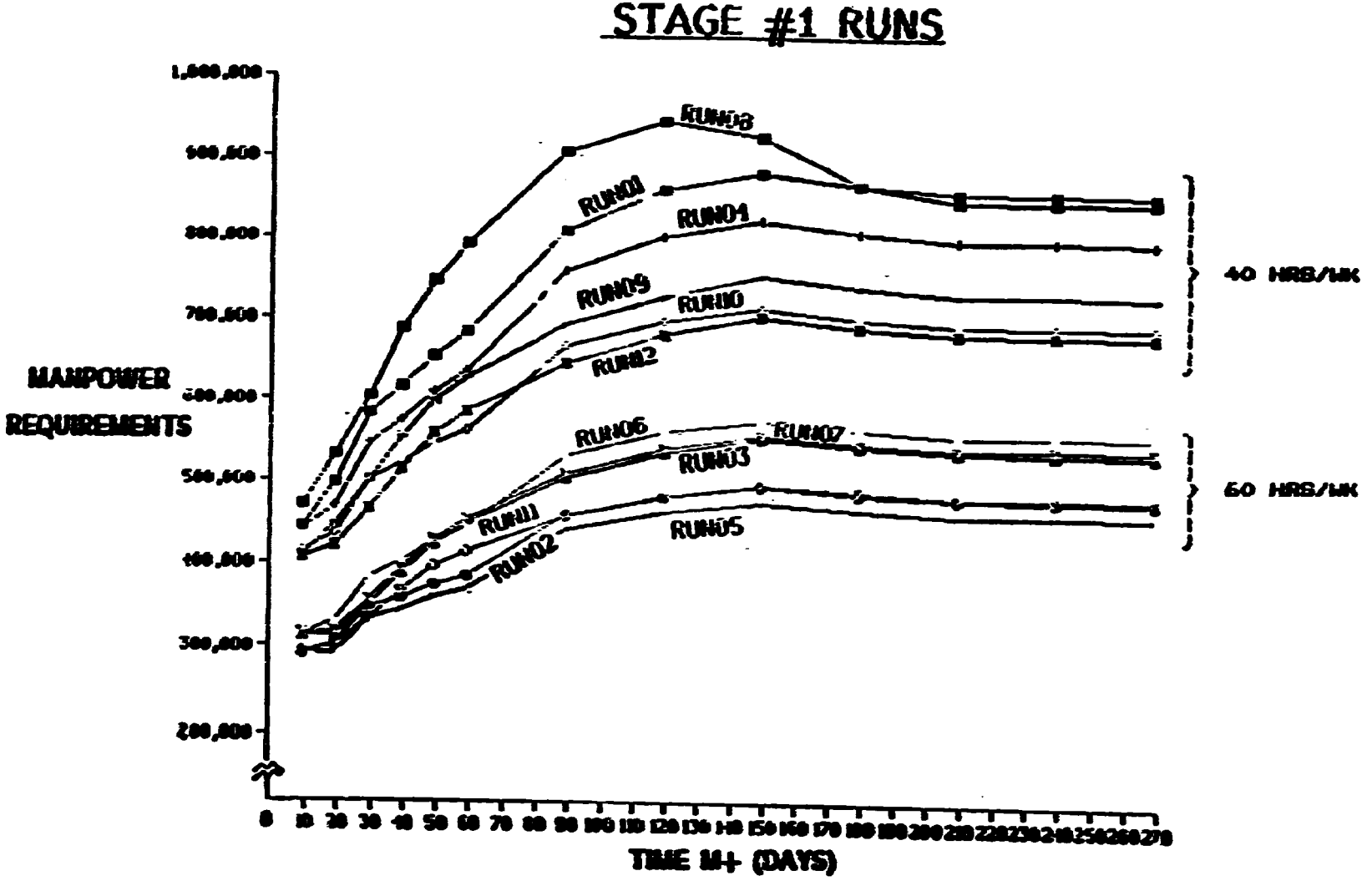

STAGE #1 RUNS
MANPOWER REQUIREMENTS
TIME M+ (DAYS)
RUN08
RUN01
RUN04
RUN09
RUN10
RUN12
RUN06
RUN07
RUN03
RUN11
RUN05
RUN02
40 HRS/WK
60 HRS/WK

Figure 1

FACTOR MAIN EFFECTS AGGREGATED OVER ALL AFDS

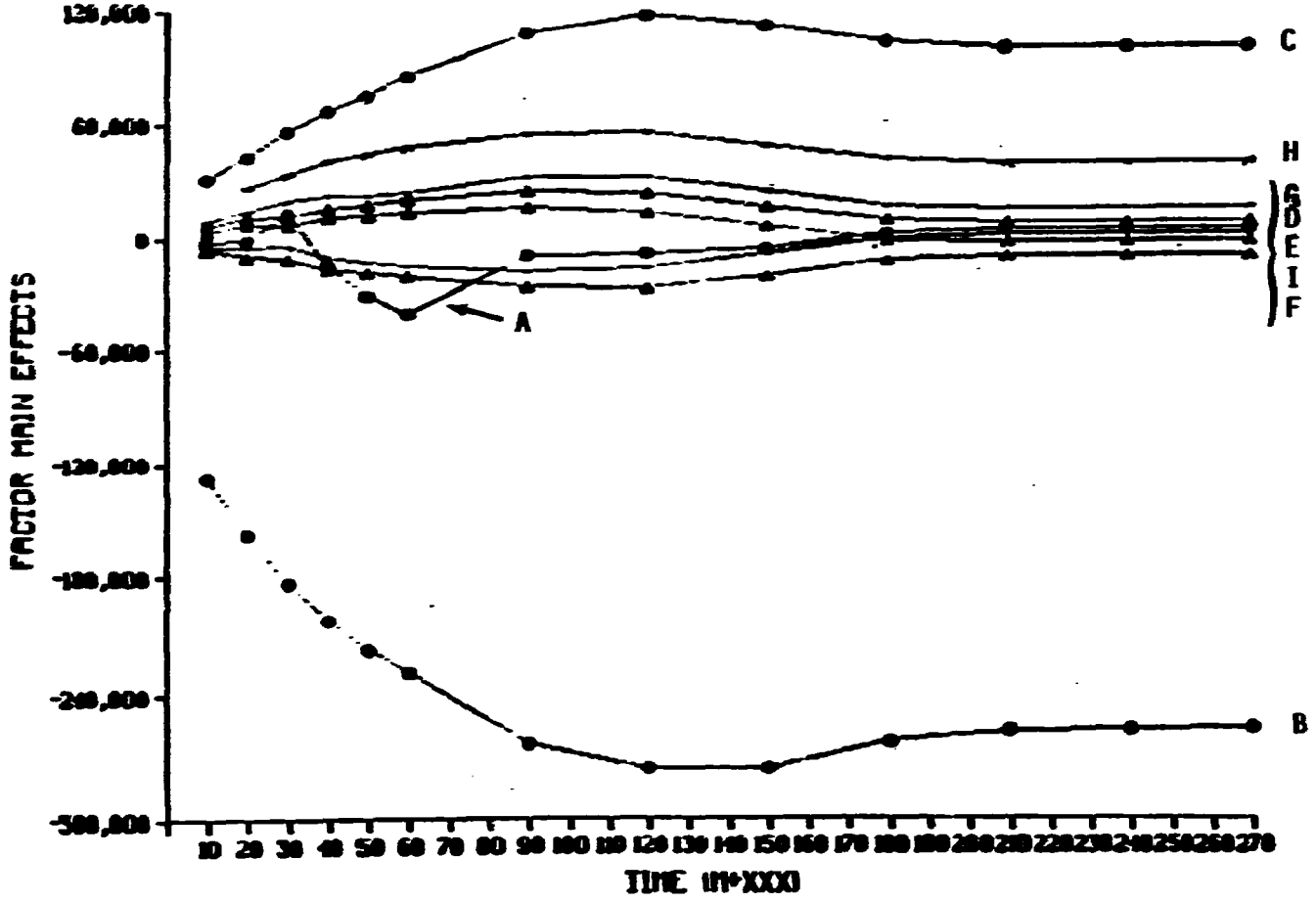


Figure 2

MANPOWER REQUIREMENTS – LEVEL MEANS

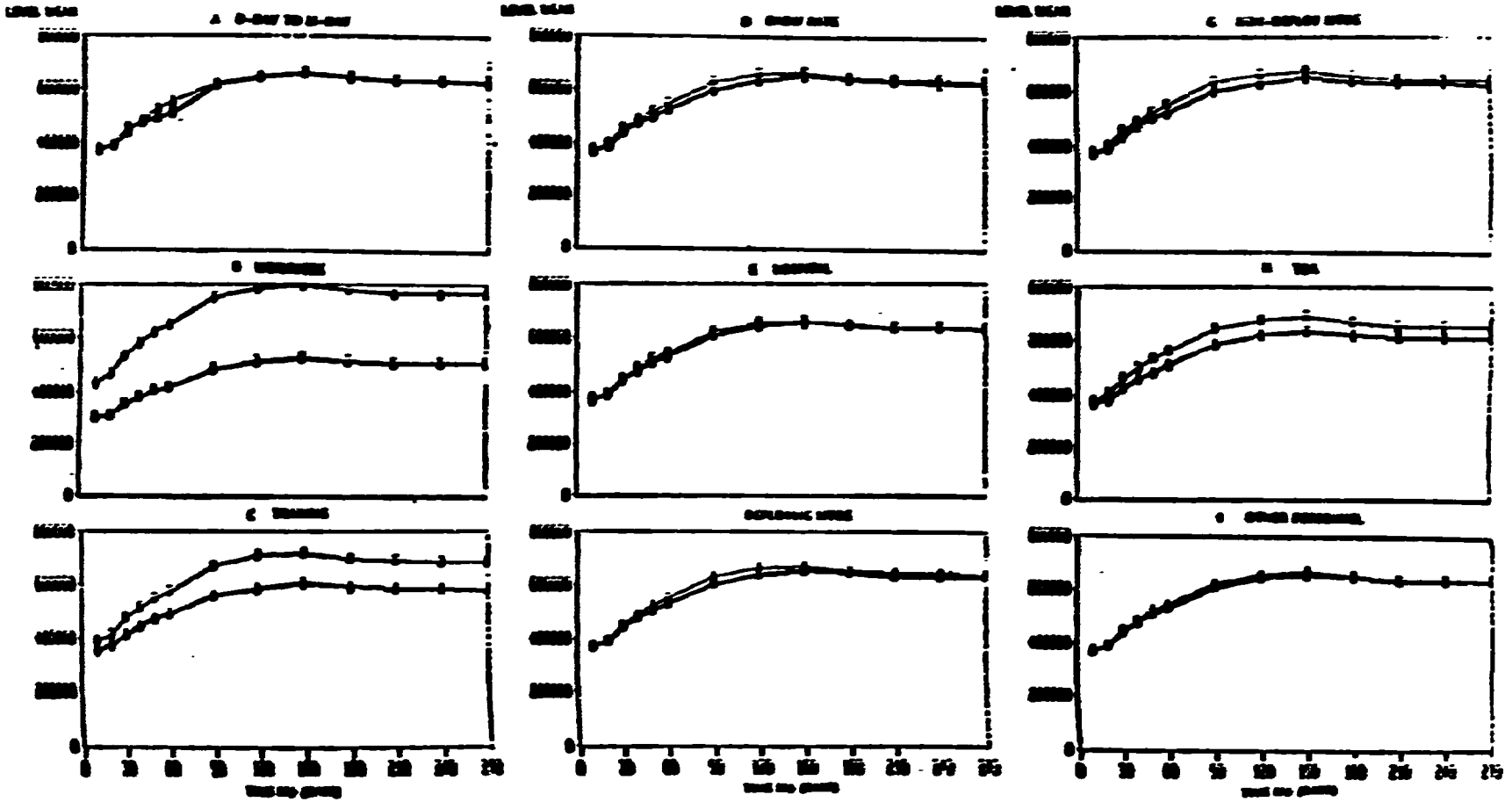


Figure 3

Table 7
Packages in the II Stage Design

Factors	Description
A	M-Day to D-Day
C1	Training load
C2	Training equipment
H1	TDA fill
H2	TDA equipment

C1 and C2 are the elements of the vector input C of the I stage design. Likewise, H1 and H2 are the components of the vector H of the I stage. At the second stage, all paramenters are scalars. The two values of the parameters at this stage are chosen within the range of their values at the first stage.

The same method of ANOVA is used as in the first stage. A sample ANOVA (for run 13) is shown in Table 8. The response variable is the manpower requirements on M + 270 day, i.e., 270 days after mobilization day. Sensitivity of a factor is measured by its contributions to the total sum of squares. The overall 'fit' is measured by 'R^2' as given below.

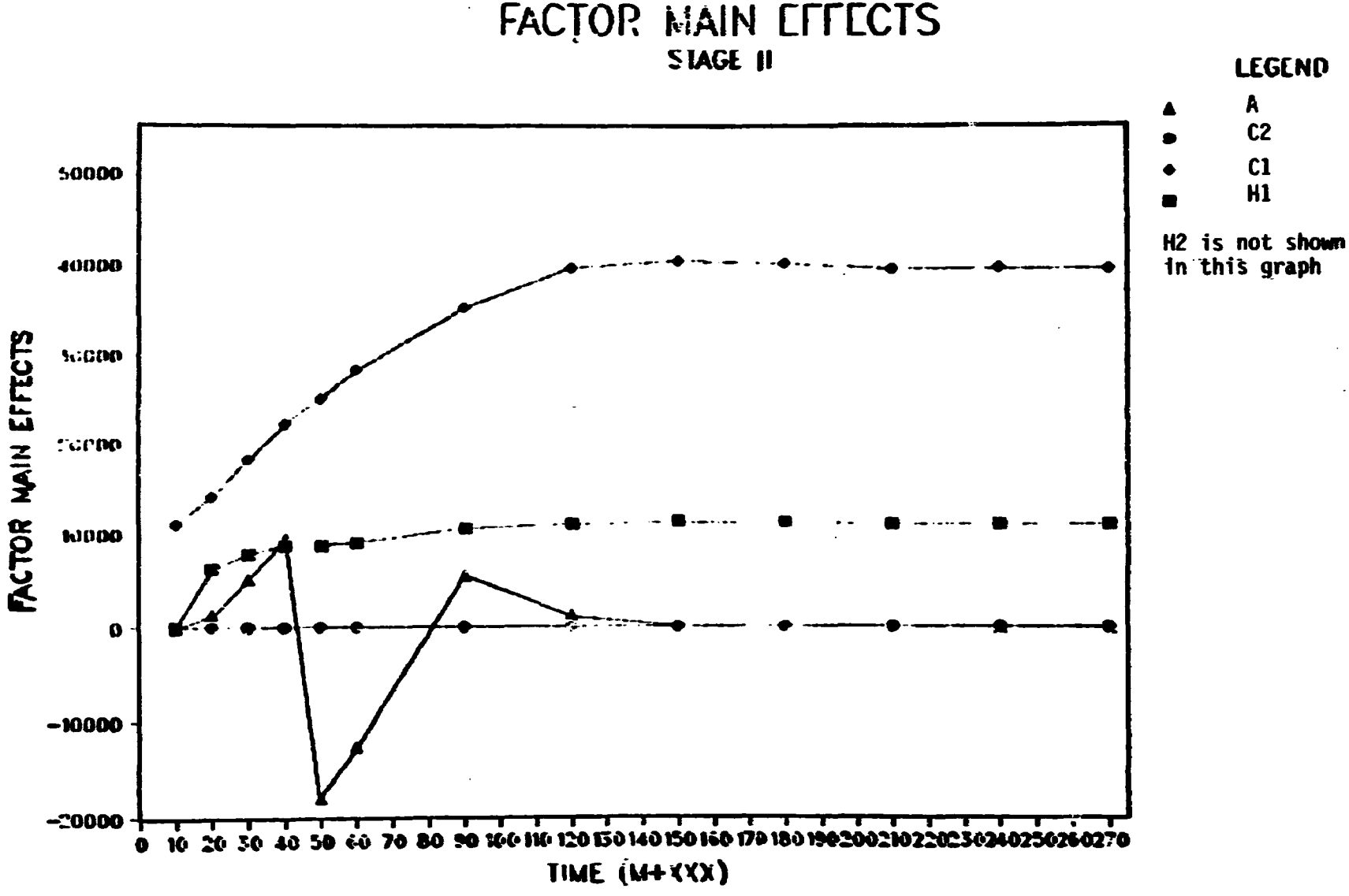

FACTOR MAIN EFFECTS
STAGE II
LEGEND
A
C2
C1
H1
H2 is not shown in this graph
FACTOR MAIN EFFECTS
50000
40000
0
-10000
-20000
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270
TIME (M+XX)

Figure 4

Table 8
ANOVA For Run 13

Source of variation	Sum of squares
A	4.000
H1	495952900.000
H2	0.250
C1	6272402402.250
C2	0.250
Explained	6768355306.750
Residual	119041.000
Total	6768474347.750

$R^2 = 93\%$

The explanation of response by the input factors are quite satisfactory with H1 and C1 being most important factors. The impact of A, H2 and C2 are negligible. Now we have 13 x 16 = 208 ANOVAs. Figure 4 shows the time series due to each of the 5 factors. Effect due to C1 is dominant, followed by H1. Effect due to A is significant up to M + 120 days, after that its impact on the response diminishes. Factors C2 and H2 are negligible.

4-3 Summary. We have summarized the data from the first stage design using regression equations. Only half the runs (B = +) from Table 2 have been utilized in deriving these equations in order to compare these results with those of the second stage design (Table 4). The regression equations and their R^2 values are given below. The dependent variable y is the manpower requirements, the independent variables are A, C1 and H1. Only the data for time phases from the mobilization day (M-Day) to 90 days after it (M + 90) are shown.

For M + 10 $y = 315567 - 3.4A + 52198C1 - 5756\,H1$

$R^2 = 99\%$

For M + 20 $y = 249976 + 77.1A + 66016C1 + 56508\,H1$

$R^2 = 97\%$

For M + 30 $y = 248656.7 + 479.1A + 869[illegible]C1 + 73635\ H1$

$R^2 = 96\%$

For M + 40 $y = 255859 - 216.7A + 104387.5C1 + 85470\ H1$

$R^2 = 98\%$

For M + 50 $y = 265077 - 644.9A + 121675.5C1 + 92054\ H1$

$R^2 = 99\%$

For M + 60 $y = 261767.3 - 882.6A + 142257C1 + 96510\ H1$

$R^2 = 99\%$

For M + 90 $y = 278884.7 + 135.6A + 173240.5C1 + 96904\ H1$

$R^2 = 99\%$

We plan to use these results along with the second stage data to apply response surface methodology for more refined predictive equations.

REFERENCES

[1] Box, G.E.P., Hunter, W.G., and Hunter, S.: *Statistics for Experimenters*, Marcel Dekker, Inc., New York (1975) , p. 431

[2] USACAA Study Report "MOBILIZATION BASE REQUIREMENTS MODEL (MOBREM) STUDY PHASES I-V," Bethesda, Maryland (1975), Executive Summary

[3] *ibid*, p. 3-1

[4] Plackett, R.L., and Burman, J.P.: "The Design of Optimum Multifactional Experiments," *Biometrika*, Vol. 33 (1946), p. 305

[5] Kleijnen, J. P.C.: *Statistical Techniques in Simulation*, Part II, Marcel Dekker, Inc., New York (1975), p. 332

[6] *op. cit.* [1], p. 421

[7] Mendenhall, W.: *Introduction to Linear Models and the Design and Analysis of Experiments*, Wadsworth Publishing Company, Inc., Belmont, California (1968), p. 98

[8] *ibid*, p. 174

ESTIMATION OF VARIANCE COMPONENTS AND MODEL-BASED DIAGNOSTICS IN A REPEATED MEASURES DESIGN

Jock O. Grynovicki
U.S. Army Laboratory Command
Human Engineering Laboratory
Aberdeen Proving Ground, Maryland
21005-5001

J. W. Green
Department of Mathematics
University of Delaware
Newark, Delaware
19711

ABSTRACT

The traditional univariate analysis of the repeated measures design is obtained by treating subjects and their associated interactions as random effects. This analysis requires that certain variances and covariances of the dependent variable at various combinations of within-subject factors be equal. Instability of the variance and covariance components may mask significant effects and compel the researcher to utilize a less powerful multivariate technique.

This paper illustrates the use of a recently developed class of unbiased variance component estimators and their associated diagnostics for examining the data and the model assumptions. A comprehensive example is given for the case of a three-way design with two factors repeated.

I. INTRODUCTION

Repeated measures designs are one of the most frequently utilized classes of designs in Army Research and Development. These designs offer a reduction in the error variance due to the removal of an individual's variability, are efficient, and require fewer subjects to achieve the same power of the F test as completely random or block designs.

This class of designs, sometimes referred to as within-subject designs, obtain their name from the fact that one or more factors of the design are manipulated in such a way that each subject receives all levels of the within subject factor. The advantage of this approach is that subjects act as their own control in their responsiveness to the various experimental treatments. On the other hand, this type of design introduces intercorrelations among the means on which the test of within subject main effects and interactions are based.

Due to this intercorrelation, three separate approaches have been proposed in the literature. The first, the univariate analysis of the repeated measures design is obtained by treating subjects as a random effect. The linear model employed is called a mixed effects model, and the resulting

analysis is a mixed model analysis of the repeated measures design. The standard mixed model assumes certain variances and covariances of responses are invariant across the experiment. For example, in a three-factor factorial model with Factors 1 and 3 fixed and subjects (or Factor 2) random, a standard assumption is that the covariance, θ_{12}, of responses at the same level of Factor 1 and on the same subject (i.e., level of Factor 2) but at different levels of Factor 3, is invariant across all subjects, all levels of Factor 1 and all combinations of distinct levels of Factor 3. More generally, if θ_t is the covariance between observations at the same levels of Factors indexed by t and at different levels of the other factors, then standard mixed models assume θ_t is invariant across all levels of the factors indexed by t and across all combinations of distinct levels of the other factors. This assumption is referred to in the literature as compound symmetry. Huynh and Feldt (1970) have shown this assumption to be a sufficient condition.

In the second approach, the multivariate method, the responses of a subject are treated as a k-dimensional response vector. It is worth noting that this approach is not as powerful as the univariate approach if the assumption of compound symmetry is accepted.

Thirdly, a degree of freedom adjustment initially proposed for use by Greenhouse and Geisser (1959) is used to adjust the numerator and denominator degrees of freedom of the ratio. Huynh and Feldt (1970) have shown this adjustment to be too conservative.

Difficulty in interpretation can occur when several dependent measures are made for each experimental treatment and the assumption of compound symmetry is rejected. This situation can result in a lack of degrees of freedom and power since the response matrix, which is a multiple of dependent variables and the number of unique within subject factor treatment combinations, can equal or exceed the total number of subjects. In the multivariate context, this can result in the degrees of freedom parameter being very small.

Since it is common and necessary to record, evaluate and analyze numerous measurements during developmental testing and human factors evaluation of weapon systems and equipment, alternative approaches to assessing the effect of treatment conditions on the response measurements need to be explored.

This paper introduces and demonstrates the use of unbiased, efficient variance component estimators and their associated diagnostics in analyzing the repeated measures design.

II. GENERAL VARIANCE COMPONENT ESTIMATES AND DIAGNOSTICS METHODOLOGY

The problem of estimating variance components in random and mixed models has been of interest to researchers for years as pointed out by Green and Hocking (1988). However, over the last few years, new closed form expressions for the estimators of variance components have been developed, based on the equivalence shown in Green (1985, 1987); Hocking, Bremer and Green (1987); and

Hocking (1985) of the variance component estimation problem to the problem of estimating the covariances, θ_t between appropriately related observations. In addition, these estimators have been shown to provide information which will be useful in diagnosing problems and suggest simple graphical procedures for examining the influence of the treatment levels.

To introduce this general methodology, this paper will only consider three factor repeated measures design with factors one and three repeated as shown in Table 1. The number of levels of factor (i) is designated by a_i. Subjects are designated factor two. Factors one and three are the within subject fixed factors. The traditional univariate repeated measures model with subject and subject interactions considered random is

$$Y(ijkm) = M + A(i) + S(j) + AS(ij) + B(k) + AB(ik) + SB(jk) + ABS(ijk) + E(ijkm)$$

where M is the overall mean, A(i) is the effect of level i of treatment or factor A, S(j) is the effect of subject j, AS(ij), is the effect of level ij of treatment combination AS, B(k) is the effect of level k of factor B, AB(ik) is the effect of the AB treatment combination at level ik, SB(jk) is the effect of treatment combination SB at level (jk), ABS(ijk) is the effect of level ijk of treatment combination ABS, and E(ijkm) is the random error. For the traditional univariate approach, it is assumed that A(i), B(k), AB(ik), and M are fixed and S(j), AS(ij), SB(jk), ABS(ijk), E(ijkm) are zero mean, independent normal random variables with variances ϕ_2, ϕ_{12}, ϕ_{23}, ϕ_{123}, and ϕ_0 respectively. While the variables are independent, the responses are correlated with the covariance structure found in Figure 1.

This covariance structure in Figure 1 suggests an alternative approach to the linear model first proposed in Hocking (1983) and extended and developed in Green (1985) to several classes of linear models. This approach relaxes the requirement that the variance components be positive. Thus, the classical model is replaced by specifying the response vector as normal with covariance matrix as given in Figure 1 and mean vector determined from the expectation of Y.

The only restriction on the covariance matrix is that it be positive definite. This requirement is weaker than the classical requirement that the ϕ_2 be positive. An in-depth development of this alternative model can be found in Hocking (1985).

The covariance, θ_t, is between observations at the same level of factors indexed by t and different levels of all other factors in the model. This suggests examining the corresponding sample covariances. These sample covariances, or averages thereof, yield the estimators of the θ_t. Sample covariances yielding estimators of θ_2 and θ_{12} are given in Figure 2. Similarly, θ_{23} is analogous to the θ_{12} estimator with subscript three replacing one. For example, from Figure 2 one recognizes the θ_2 estimator as the average of $a_{13}r_{13}$ equal expectation sample covariances corresponding to all combinations of $i \neq i^*$, $k \neq k^*$. Here r_i is the level of Factor i minus one. Similarly, θ_{12} is the average of $a_{13}r_3$ equal expectation sample covariances corresponding to all combinations of i and $k \neq k^*$.

TABLE I
THREE FACTOR REPEATED MEASURES DESIGN

FACTOR I

	1				2				3				...				a_1			
FACTOR III																				
SUBJECTS	1	2	3	a_3	1	2	3	a_3	1	2	3	a_3	1	2	3	a_3	1	2	3	a_3
1																				
2																				
3																				
4																				
.																				
.																				
.																				
a_2																				

COVARIANCE STRUCTURE
2 WITHIN SUBJECT FACTORS

$$\text{COV}\ (Y(I,J,k,m), Y(I^*,J^*,k^*,m^*)) =$$

$$\theta_2 = \phi_2 \quad \text{if} \quad I \neq I^*,\ J = J^*\ k \neq k^*$$
$$\theta_{12} = \phi_2 + \phi_{12} \quad \text{if} \quad I = I^*,\ J = J^* \quad k \neq k^*$$
$$\theta_{23} = \phi_2 + \phi_{23} \quad \text{if} \quad I \neq I^*,\ J = J^*\ k = k^*$$
$$\theta_{123} = \phi_2 + \phi_{12} + \phi_{23} + \phi_{123} \quad \text{if} \quad I = {}^* \ IJ = J^*\ k = k^*\ m^* \neq m$$

$$\phi_o + \theta_{123} \qquad IJkm = I^*J^*k^*m^*$$

Figure 1: Covariance structure of three repeated measures design (Subjects random)

VARIANCE COMPONENT ESTIMATES

$$\hat{\theta}_2 = \frac{1}{r_2\ a_{13}\ r_{13}} \sum_{Ik \neq I^*k^*} \sum_{J} (\overline{yIJk} - \overline{yI.k.})\ (\overline{yI^*Jk^*} - \overline{yI^*.J^*.})$$

$$\hat{\theta}_{12} = \frac{1}{a_{13}\ r_3} \sum_{k} \frac{1}{r_2} \sum_{IJ} (\overline{yIJk.} - \overline{yI.k.})(\overline{yIJk^*.} - \overline{yI.k^*.})$$

Figure 2: Variance component estimates for Θ_2 and Θ_{12}

These covariances are unbiased and contain the diagnostic power. By plotting these covariances (diagnostics) in table form, one obtains an indication of the stability of the estimate and of suspect estimates.

In general, one looks for various characteristics and trends. For example, (1) unusually large or small diagonal entries indicate abnormal variability in the cell means for this level of the factor under investigation, (2) special patterns in the off-diagonal elements such as a particular column or row having the majority of its entries higher or lower than associative rows or columns, indicate one or more cell means may contain extreme outliers, and (3) large fluctuations in the off-diagonal entries reflect high variability is the data.

Following the examination of the diagnostics, plots of treatment i vs. treatment i* cell-means, where abnormal diagnostics have been identified, are recommended. This will help the researcher identify the treatment cells responsible for extra large or small variance component estimates. Finally, the diagnostic procedure should conclude with an examination of the data in the identified cells.

III. REPEATED MEASURE DESIGN

To illustrate these diagnostic procedures, data from a repeated measures design carried out by Malkin and Christ (1987) will be used.

A. Objective

The objective of the experiment was to conduct a laboratory flight simulation to compare a cockpit keyboard, a thumb-controlled switch, and a connected-word voice recognizer for data entry of navigation map coordinate sets when (1) the entry of Universal Transverse Mercator (UTM) coordinate sets is the sole task performed (No Flight) and (2) the entry of UTM coordinate sets is performed concurrently with controlling a helicopter simulator while flying a computer-generated external scene (Flight). For this paper, the difference among the three methods of data entry for response and input time will be evaluated for both the Flight and No Flight conditions. The original paper also investigated error. However, no practical or statistical difference was found for subject error in regard to any of the experimental factors.

B. Methodology

Data were collected using 12 Army aviators assigned to Aberdeen Proving Ground, Maryland as the experimental units.

The Aviation and Air Defense Division, Human Engineering Laboratory's (HEL's) flight simulator was utilized for this study. The Crew Simulator

consists of a cockpit cab with advanced controls and displays and an "out-the-window" scene produced by Computer-Generated Imaging (CGI). The CGI, cockpit controls, flight simulation, displays and results were driven or recorded using two Vax computers. Training was administered to all subjects in the operation of the voice recognition system and flight simulator. For an in-depth accounting of the Apparatus and Training, the reader is referred to Malkin and Christ (1987).

C. Procedure

Each subject entered eight UTM coordinate sets for each test condition. The coordinate sets, which were selected from a scenario based on the Fulda Gap area of Germany, were located on a kneeboard attached to the subject's leg. A standardized, but different set of coordinates was used in each condition. The subject was tested in both conditions using one data entry method before proceeding to the next data entry method. The order of the test conditions were counterbalanced to control for learning.

D. Experimental Design

A 2x3x12 factorial design with repeated measures on the twelve subjects was implemented. The within subject factors were data entry methods (voice, keyboard and thumb-controlled switch) and task conditions (flight, no flight). The dependent variables were input time and response time. For illustration, the 2x3x12 repeated measures design along with input time can be found in Table 2.

E. Results

Since the response measures were highly correlated, and only 12 subjects were used, a multivariate analysis of variance was performed using the univariate repeated measures model with subjects considered a random factor. The approximate F ratios were then checked against the Greenhouse Geisser adjustment and they agreed.

The results are shown in Figure 3. For response time, subjects were able to respond significantly faster during the no-flight condition than during the flight condition. There also was a significant interaction between data entry method and task conditions. During the no-flight task condition, subjects responded significantly faster when the keyboard was used to enter data. However, during the flight task condition, subjects responded significantly faster using either voice or the thumb-controlled switch (see Figure 4).

There were significant differences among the three mean impact times for the data entry method. Subjects were also able to input data faster during the no-flight task conditions than during the flight conditions. However, there was no significant interaction between Task and Entry method (see Figure 5).

TABLE 2. METHOD BY TASK BY SUBJECT
(INPUT TIME)

	Method					
	Voice 1 Task		Keyboard 2 Task		Thumb 3 Task	
Subject	No Flight 1	Flight 2	No Flight 1	Flight 2	No Flight 1	Flight 2
1	15.8	17.8	16.9	16.8	28.5	34.3
2	23.9	49.3	9.1	13.2	25.0	35.5
3	33.0	55.9	13.6	31.6	29.7	48.8
4	15.2	27.8	11.3	16.1	24.1	43.1
5	35.9	45.0	11.9	20.7	39.2	65.2
6	49.8	36.4	11.8	23.7	36.3	49.1
7	27.2	34.9	13.9	20.6	31.7	44.7
8	20.6	20.6	10.9	24.1	35.4	37.4
9	28.92	38.7	10.5	19.9	34.7	34.6
10	27.7	23.5	10.7	15.9	34.0	43.6
11	17.9	11.7	15.4	24.1	32.6	39.0
12	23.0	16.3	13.5	33.8	38.9	70.9

MANOVA RESULTS

STATISTICAL SIGNIFICANCE OF FACTORS BY DEPENDENT MEASURES

DEPENDENT MEASURES	SUBJECT F-STATISTICS	SUBJECT SIGNIFICANCE = .05
RESPONSE TIME	8.64	*
INPUT TIME	5.57	*

DEPENDENT MEASURES	METHOD F-STATISTICS	METHOD SIGNIFICANCE = .05
RESPONSE TIME	1.07	
INPUT TIME	30.78	*

DEPENDENT MEASURES	TASK F-STATISTICS	TASK SIGNIFICANCE = .05
RESPONSE TIME	29.25	*
INPUT TIME	25.21	*

DEPENDENT MEASURES	METHOD BY TASK F-STATISTICS	METHOD BY TASK SIGNIFICANCE = .05
RESPONSE TIME	11.7	*
INPUT TIME	2.25	

DEPENDENT MEASURES	SUBJECT BY METHOD BY TASK F-STATISTICS	SUBJECT BY METHOD BY TASK SIGNIFICANCE = .05
RESPONSE TIME	1.13	
INPUT TIME	1.00	

DEPENDENT MEASURES	SUBJECT BY TASK F-STATISTICS	SUBJECT BY TASK SIGNIFICANCE = .05
RESPONSE TIME	4.09	*
INPUT TIME	2.45	*

DEPENDENT MEASURES	SUBJECT BY METHOD F-STATISTICS	SUBJECT BY METHOD SIGNIFICANCE = .05
RESPONSE TIME	3.15	*
INPUT TIME	2.09	*

Figure 3: Manova results for three factors repeated measures design with response time and input time the dependent measure.

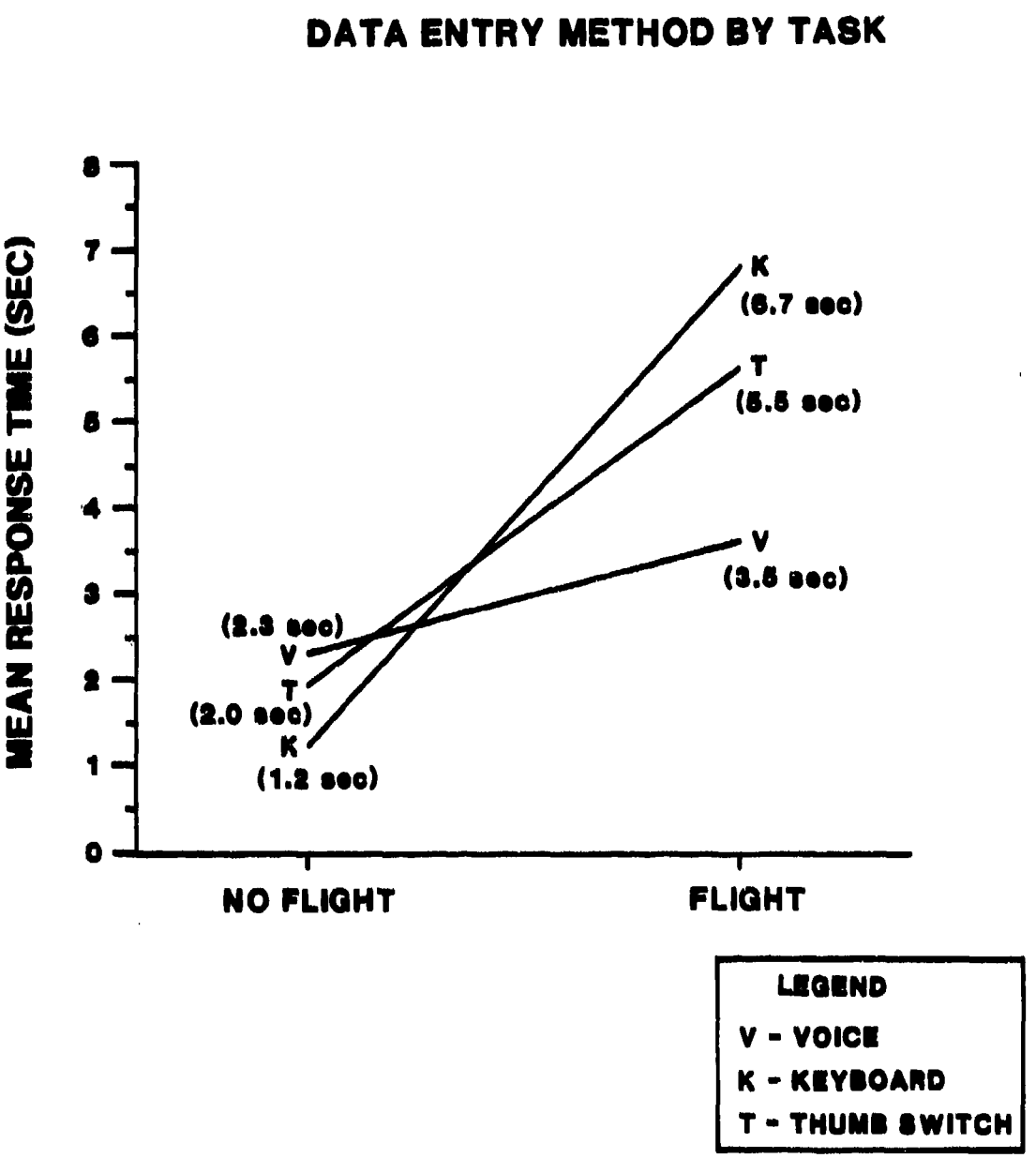


Figure 4: Data entry methods by task for response time

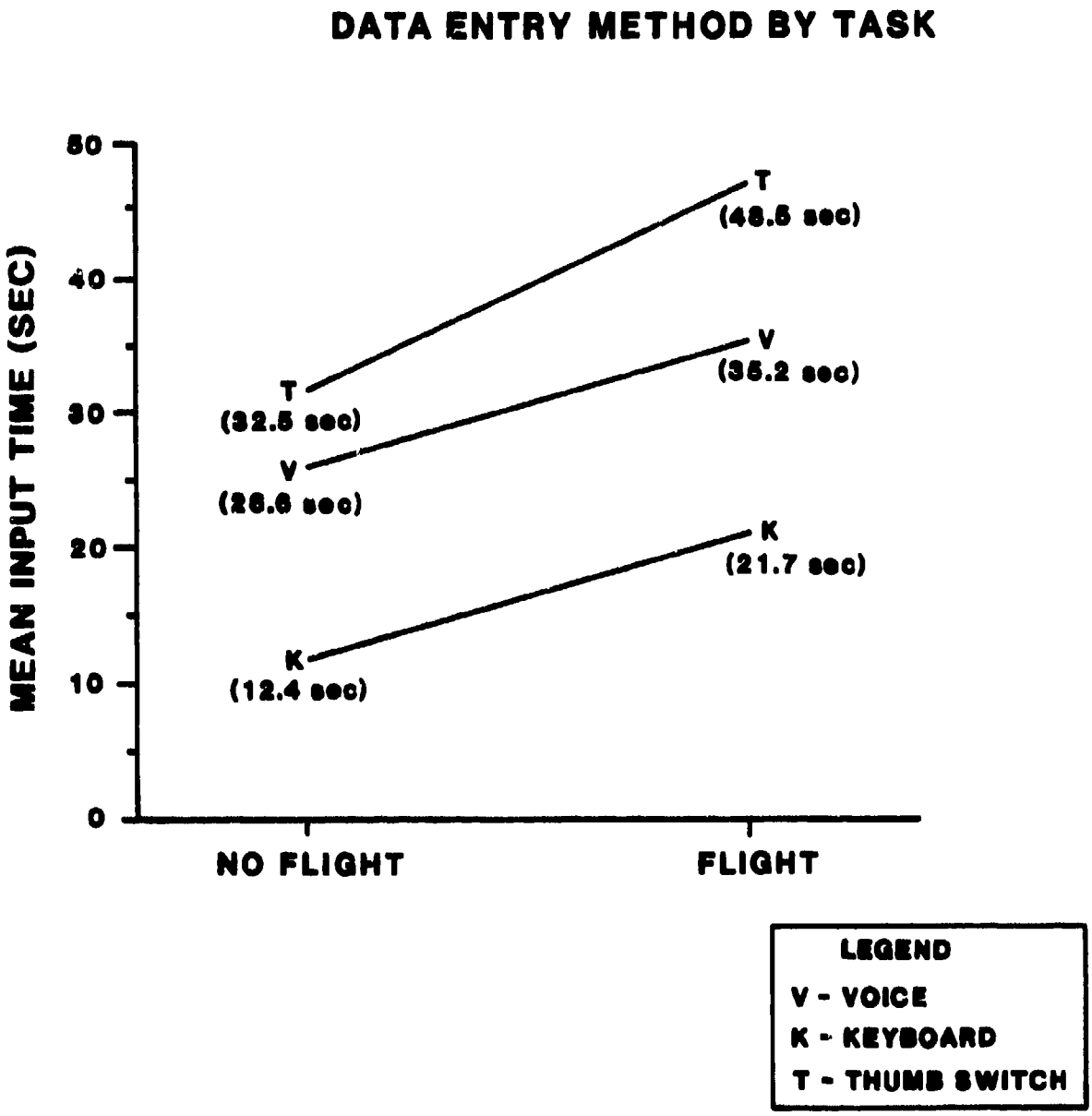


Figure 5: Data entry method by task for input time

As a final note, the input time covariances for the within-subject factors deviated extremely from the compound symmetry assumption whereas the compound symmetry assumption for response time was acceptable. Therefore, the variance component diagnostic procedure will be demonstrated for input times only.

IV. ILLUSTRATED EXAMPLE OF VARIANCE COMPONENT ESTIMATES AND DIAGNOSTICS

As previously pointed out, it is natural to estimate the covariances θ_t by corresponding sample covariances. In the balanced case, and for the Malkin, Christ data, the estimates can be obtained from the ANOVA table (see Figure 6).

For this example, $a_1 = 3$, $a_2 = 12$ and $a_3 = 2$. The estimate of θ_2 is the average of six distinct sample covariances. They can be displayed in a table such as Table 3-A. The off-diagonal elements are the sample covariances. To avoid confusion, it is worth noting that the diagonal elements are not true variances since $i \neq i^*$. An alternative and simpler display of these sample covariances can be found in Table 3-B. Again, the diagonal elements are not true variances since $k \neq k^*$.

Under the compound symmetry assumption, all elements of Table 3-A or Table 3-B should be approximately equal. Therefore, the diagnostics provide a illustrative procedure to check the compound symmetry assumption and identify unique treatments combinations that contribute to this assumption being violated.

In examining the θ_2 off-diagonal diagnostics of Table 3-A, the covariances Keyboard No Flight vs. Voice Flight (-13.81) and Thumb No Flight vs. Voice Flight (-12.47) are small when compared to the other off-diagonal entries in the Table. In addition, Thumb Flight vs. Voice No Flight (40.78) seems large in comparison. This large fluctuation indicates high variability in the data.

The diagonal entries of Table 3-A indicates the covariances at the same Task level but different Input levels. The large diagonal entry (43.26), representing the covariance of Thumb Flight vs. Keyboard Flight, indicates instability and variability in the cell means making up this covariance. Referring to Table 1, the reader can see that the cell means for Keyboard, Flight and Thumb Flight are larger and more unstable than the other Method Task treatment conditions.

This suggests further examination of the specified treatment combinations. Follow-up plots of subject mean input times by treatment combinations reflecting the large or small covariances are shown in Figures 7 through 9.

Examinatin of these plots revealed that subjects (3, 5, 6 and 12) input time contributed to the extremely high or low covariances.

ANOVA

SOURCE	dF	E M S
METHOD	2	$\theta_o + n\Phi_{123} + na_3\Phi_{12} + na_2a_3M_1$
TASK	1	$\theta_o + n\Phi_{123} + na_1\Phi_{23} + na_1a_2T_3$
METH x TASK	2	$\theta_o + n\Phi_{123} + na_2MT_{13}$
SUB	11	$\theta_o + na_1a_3\Phi_2 + na_3\Phi_{12} + n\Phi_{123}$
SUB x METHOD	22	$\theta_o + na_3\Phi_{12} + n\Phi_{123}$
SUB x TASK	11	$\theta_o + na_1\Phi_{23} + n\Phi_{123}$
SUB x METHOD x TASK	22	$\theta_o + n\Phi_{123}$
ERROR	504	θ_o

Figure 6: Analysis of variance for the three way repeated measures model. Method and task are within subject factors. Subjects are considered random.

TABLE III - A

DIAGNOSTIC

INPUT TIME

θ_2

KEYBOARD		VOICE	
		NO FLIGHT	FLIGHT
	NO FLIGHT	-5.90	-13.81
	FLIGHT	13.68	-5.07

THUMB		VOICE	
		NO FLIGHT	FLIGHT
	NO FLIGHT	23.15	-12.47
	FLIGHT	40.78	10.52

THUMB		KEYBOARD	
		NO FLIGHT	FLIGHT
	NO FLIGHT	0.19	16.01
	FLIGHT	1.88	43.26

TABLE III - B

DIAGNOSTIC

INPUT TIME

Θ_2

NO FLIGHT

FLIGHT	VOICE	KEYBOARD	THUMB
VOICE	76.20	-13.81	-12.47
KEYBOARD	13.68	4.80	16.01
THUMB	40.78	1.88	35.10

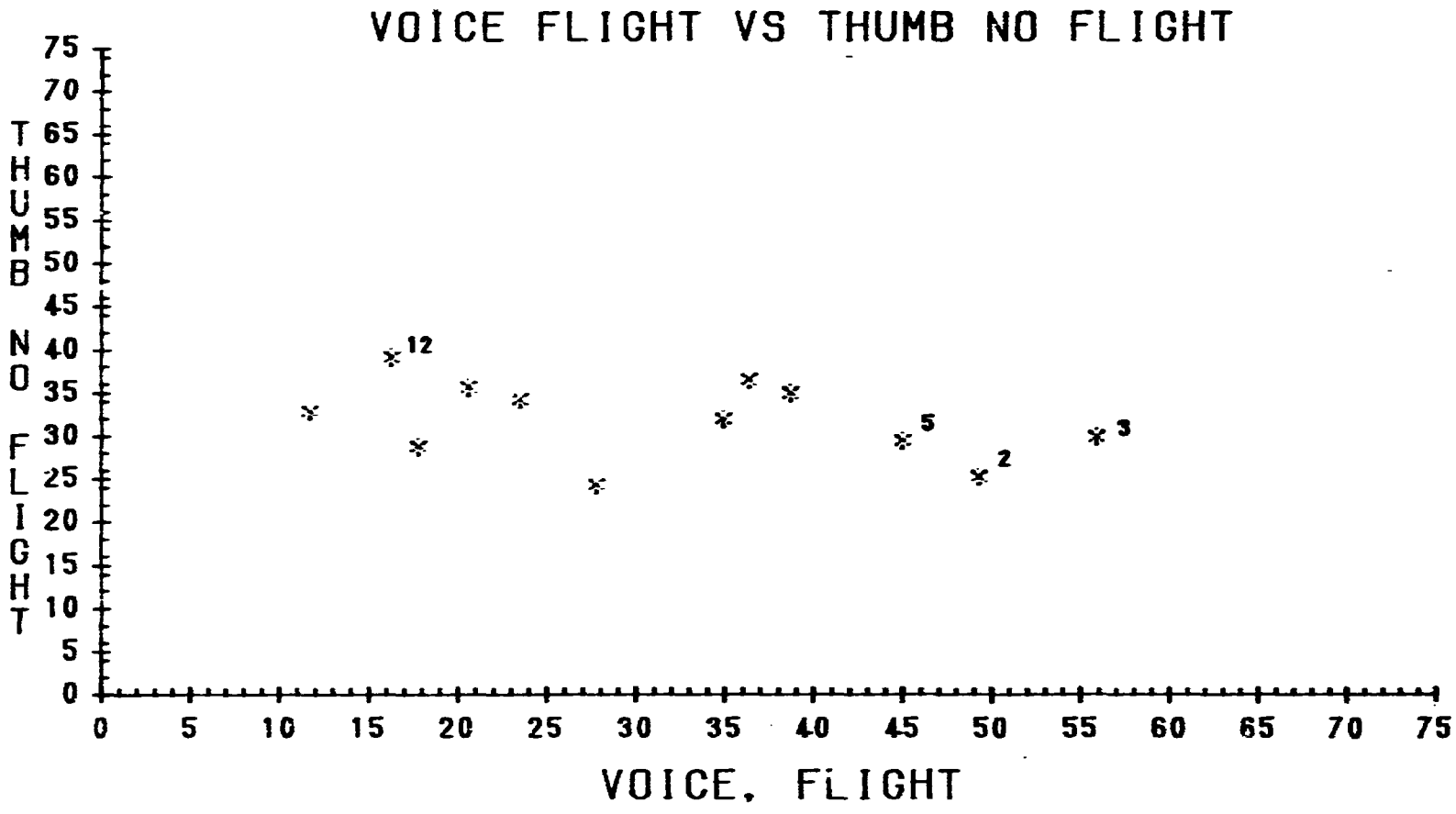


Figure 7: Malkin and Christ Data: Subject cell means of Thumb No Flight VS Voice Flight. Plotting Symbol: numbers represent a distinct subject.

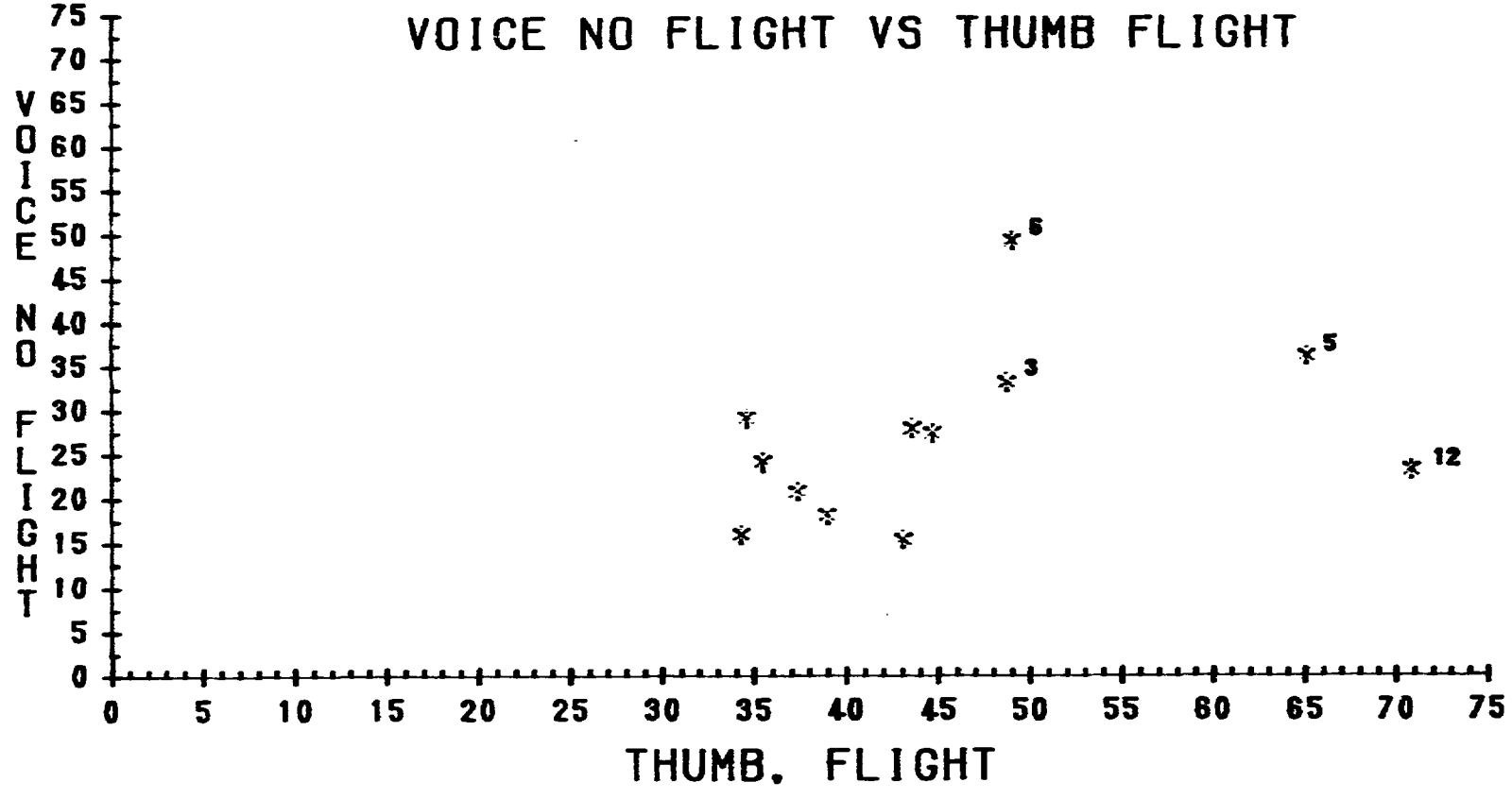


Figure 8: Malkin and Christ DATA: Subject cell means of Voice No Flight VS Thumb Flight. Plotting Symbol: numbers represent a distinct subject.

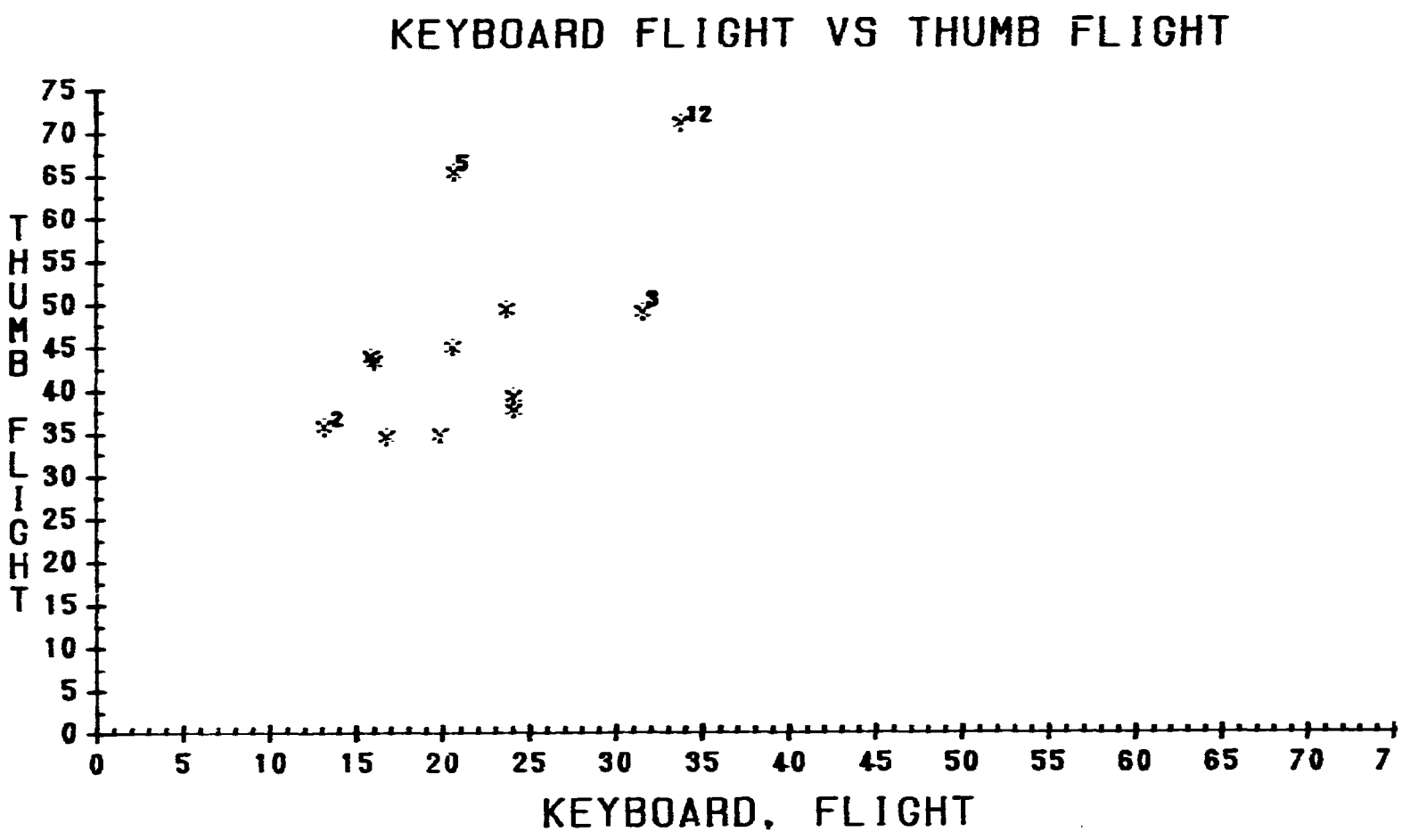


Figure 9: Meikle and Christ DATA: Subject cell means of Keyboard Flight VS Thumb Flight. Plotting Symbol: numbers represent a distinct subject.

The diagnostic plots for θ_{12} and θ_{23} are shown in Table 4. For θ_{12}, the plot consists of covariances based on the same level of Subject and Method, but different levels of Task. The diagnostic plot revealed a spurious covariance component of 76.2 for Voice No Flight vs. Voice Flight. A follow-up plot (Figure 10) indicated that subjects (3, 5 and 6) input times contributed to this large covariance.

Similarly, the diagnostic plot for θ_{23}, revealed large spurious covariances at treatment combinations Voice No Flight vs. Thumb No Flight (23.1) and Keyboard Flight vs. Thumb Flight (43.2).

It is worth noting that this diagnostic plot contains covariances based on the same subject and Task levels but different Methods.

Follow-up plots (Figures 11, 12) for both covariances revealed that subjects (3, 5, 6 and 12) input time were contributing to one or both large covariance components.

Identifying what seemed to be a dichotomous population of subjects, a review of subject records were undertaken to attempt to explain the reason subjects 3, 5, 6 and 12 seemed to respond differently from the rest of the subjects. A review of the records indicated that, in general, these pilots were older (over 42 as compared to under 38), had a higher military rank, and had spent as much time or more flying fixed wing or rotary wing aircraft, with recent flying experience concentrated on fixed wing. Based on subjective input from experienced pilots, differences between the aircraft in regard to instrumentation and flying procedures could certainly account for the difference in input times between fixed wing and rotary wing pilots.

A recalculation of the diagnostics with subjects 3, 5, 6 and 12 removed revealed covariances that were more stable. In addition, in grouping the subjects into Fixed Wing and Rotary Wing categories and reanalyzing the data, the assumption of compound symmetry was accepted. Mauchly's criteria, which is used to check this assumption, was found not to be significant at the .01 level.

This information was made available to the Aviation and Air Defense Division of the HEL so that this additional source of variability could be controlled for future experiments.

V. CONCLUSIONS

The variance component estimates and associated diagnostic procedures have been shown to be computationally and intuitively simple. All calculations can be obtained using standard statistical packages such as SPSSX, SAS, or BMDP.

The diagnostic procedures have been demonstrated to be effective in checking underlying assumption (compound symmetry) of the repeated measures model, and useful in identifying probable causes for the violation of these

TABLE IV

DIAGNOSTIC

INPUT TIME

θ_{12}

METHOD (i)

1*=1	VOICE TASK (k)		KEYBOARD TASK (k)		THUMB TASK (k)	
	NO FLIGHT (1)	FLIGHT (2)	NO FLIGHT (1)	FLIGHT (2)	NO FLIGHT (1)	FLIGHT (2)
NO FLIGHT	95.3	76.2	5.1	4.8	24.09	35.1
FLIGHT	76.2	19.9	4.8	38.7	35.1	138.1

θ_{23}

TASK (k)

	NO FLIGHT METHOD (i)			FLIGHT METHOD (i)		
	VOICE	KEYBOARD	THUMB	VOICE	KEYBOARD	THUMB
VOICE	95.3	-5.9	23.1	199.0	-5.0	10.5
KEYBOARD	-5.9	5.1	0.19	-5.0	38.7	43.2
THUMB	23.1	0.19	24.1	10.5	43.3	138.2

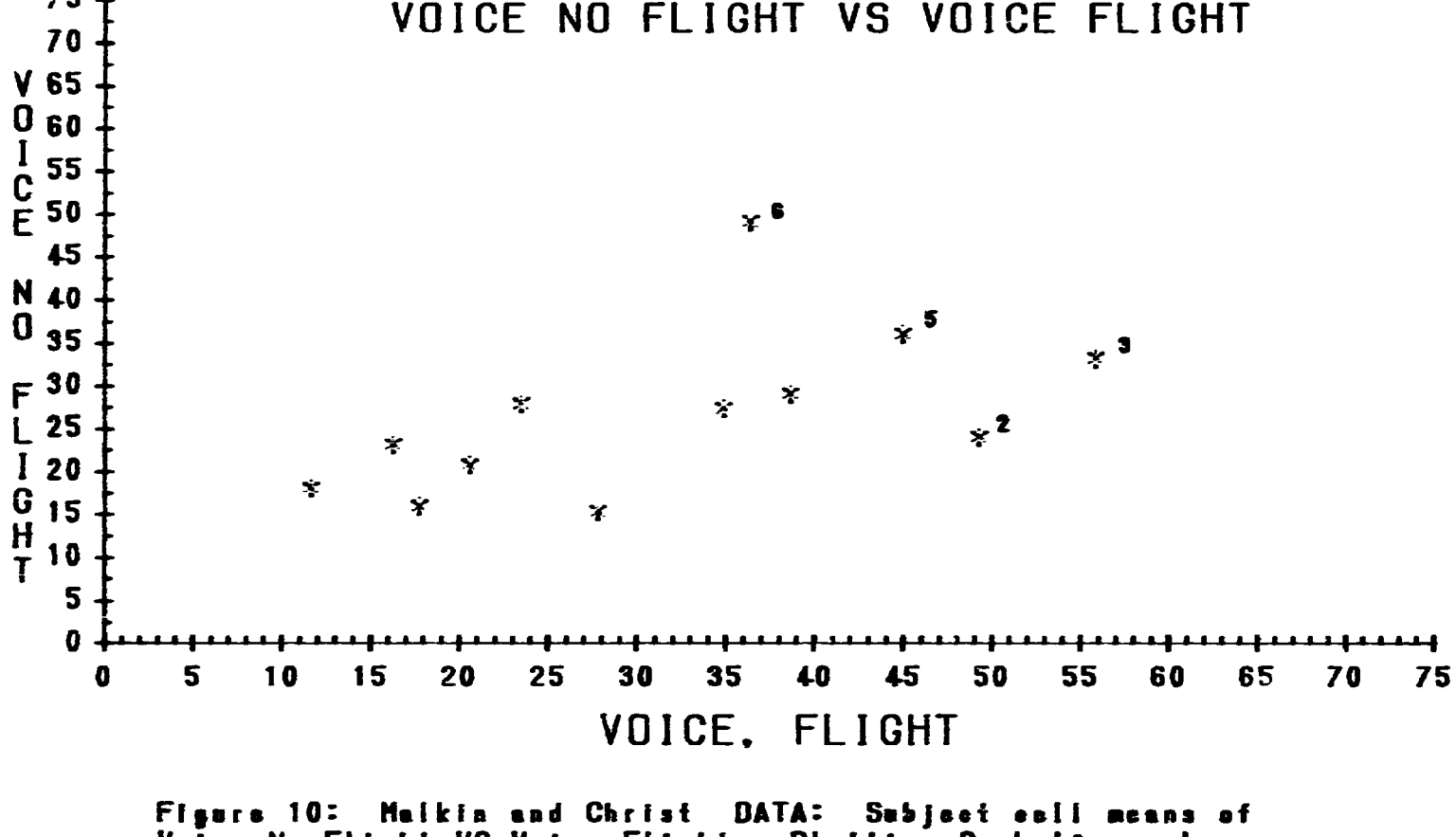


Figure 10: Malkin and Christ DATA: Subject cell means of Voice No Flight VS Voice Flight. Plotting Symbol: numbers represent a distinct subject.

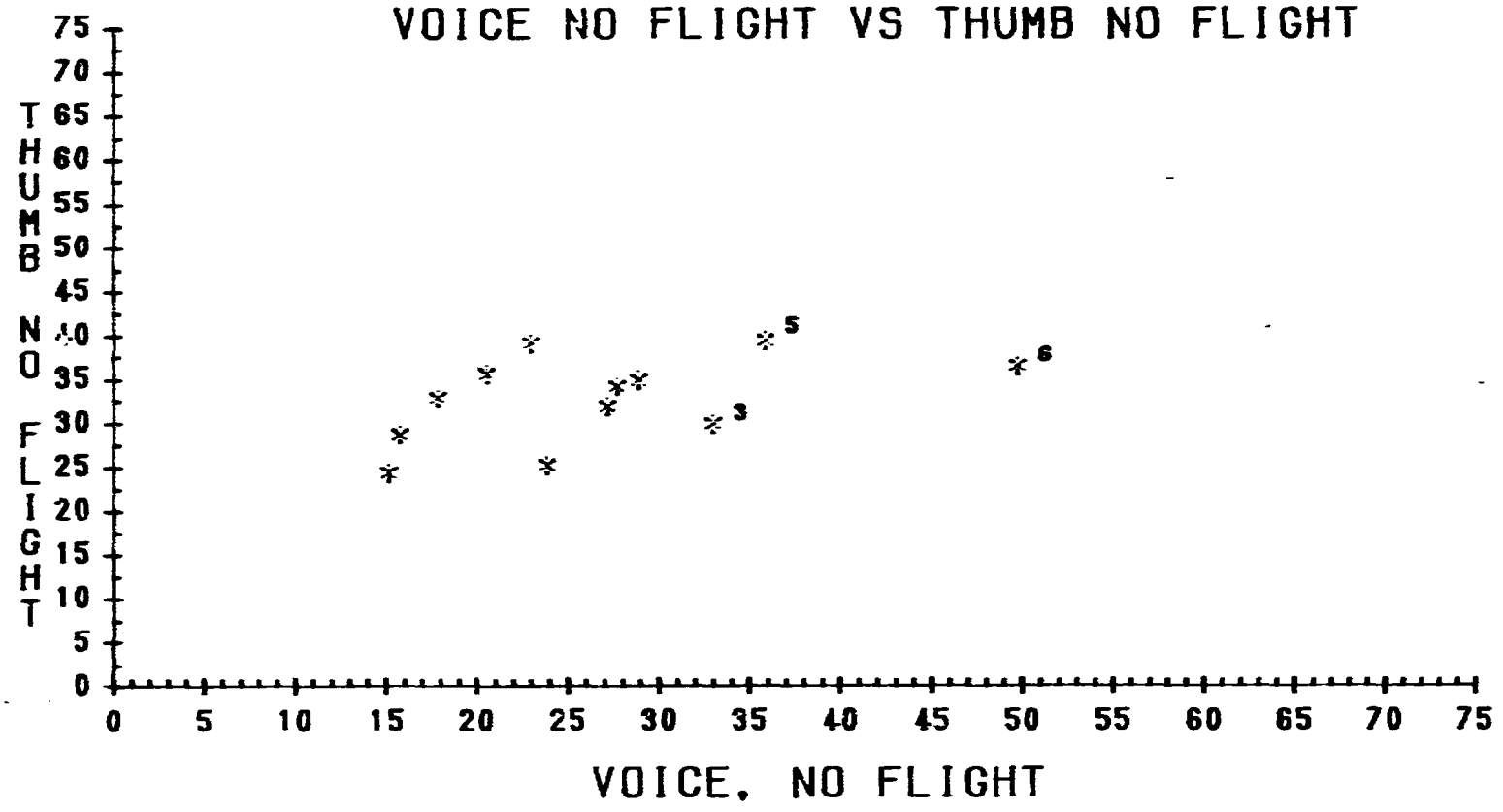


Figure 11: Malkin and Christ DATA: Subject cell means of Voice No Flight VS Thumb No Flight. Plotting Symbol: numbers represent a distinct subject.

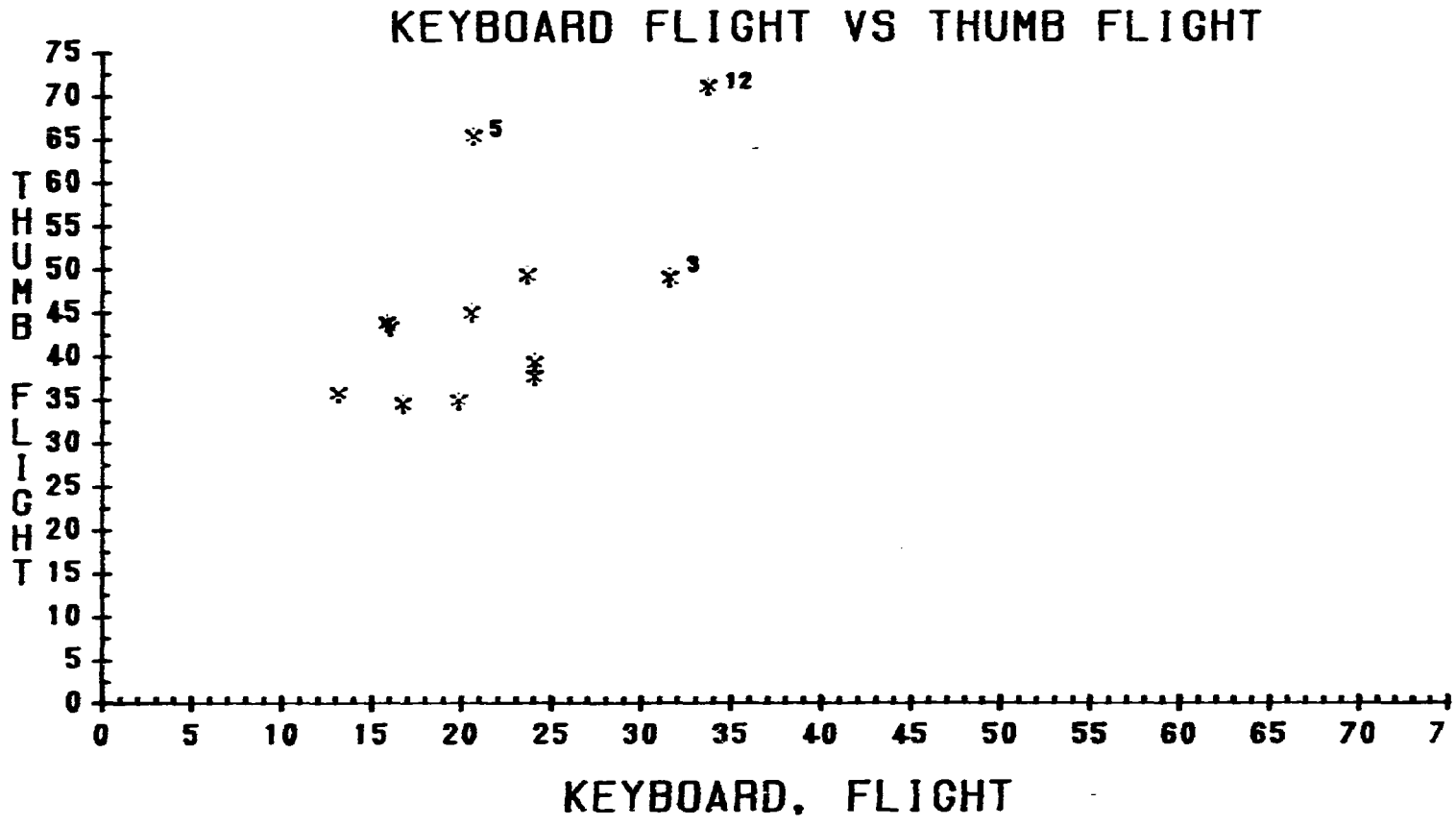


Figure 12: Malkin and Christ DATA: Subject cell means of Keyboard Flight VS Thumb Flight. Plotting Symbol: numbers represent a distinct subject.

assumptions. This provides the researcher the option of removing spurious observations, performing transformations, or controlling additional sources of variability so that the data can conform to the standard assumptions such as compound symmetry or to modifying the model. By circumventing the problems associated with the traditional univariate repeated measures analysis, these diagnostic procedures provide easier interpretation of the results and increased validity of the conclusions derived from the data. The result is a valuable statistical approach that can be applied in many areas including developmental testing and human factors evaluation of weapon systems and equipment.

REFERENCES

Green, J.W. (1985). Variance Components: Estimates and Diagnostics, Doctoral Dissertation, Texas A & M University.

Green, J.W. (1987). Diagnostic Procedure for Variance Components for Both Large and Small Designs, submitted to Technometrics.

Green, J.W. (1988). Diagnostics for Variance Components Suitable for Designs of All Sizes, submitted to Technometrics.

Green, J.W. and Hocking, R. R. (1988). Model Based Diagnostics for Variance Components In a General Mixed Linear Model, Proceedings of the Thirty-Third Conference on the Design of Experiments in Army Research, Development and Testing.

Greenhouse, S.W. and S. Geisser (1959). On Methods in the Analysis of Profile Data. Psychometrika, 24, 95-112.

Hocking, R.R. (1983). A Diagnostic Tool for Mixed Models with Applications to Negative Estimates of Variance Components, Proceedings of the SAS Users Group, New Orleans, LA, 711-716.

Hocking, R.R. (1985). The Analysis of Linear Models, Monterey, CA: Brooks-Cole.

Hocking, R.R., Bremer, R.H. and Green, J.W. (1987). Estimation of Fixed Effects and Variance Components in Mixed Factorial Models Including Model-Based Diagnostics, submitted to Technometrics.

Huynh, H. and Feldt, L.S. (1970). Conditions Under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions. JASA, 65, 1582-1589.

Malkin, F.J. and Christ, K.A. (1987). Comparison of Alphanumeric Data Entry Methods for Advanced Helicopter Cockpits (TM-14-87). Aberdeen Proving Ground, MD: U.S. Army Laboratory Command, Human Engineering Laboratory.

MODEL BASED DIAGNOSTICS FOR VARIANCE COMPONENTS IN A GENERAL MIXED LINEAR MODEL

J. W. Green
Department of Mathematical Sciences
University of Delaware
Newark, Delaware 19711

R. R. Hocking
Department of Statistics
Texas A&M University
College Station, Texas 777843

ABSTRACT

A new class of unbiased estimators is given for unbalanced mixed models which have simple, closed-form expressions. These estimators allow easy computation of variances which, when compared to minimum variance bounds, show the estimators to be highly efficient.

Based on the estimator, a diagnostic methodology is developed for assessing the effect of the data on the estimates. The source of negative estimates of variance components is often revealed, as well as other sorts of instability and problems with the model or data.

An overview of the methodology and its growing literature is given, illustrated by applications to several industrial problems. The methodology applies to all random and mixed models, regardless of the degree of imbalance or pattern of crossed and nested factors. The diagnostics flag only those features of the data which affect parameter estimates.

1. INTRODUCTION

The problem of estimating variance components in random and mixed models has become a classical research area in statistics. Review papers, such as those by Searle (1971), Harville (1977), Sahai (1979),

Sahai and Khuri (1984) and Khuri and Sahai (1984), attest to the importance of the problem and emphasize the fact that there are many aspects of the problem which remain unsolved.

It is well known that in the case of balanced data, The ANOVA estimators, or, equivalently, the restricted maximum likelihood estimators (REML), have certain optimality properties. Graybill and Hultquist (1961) showed that these estimators are uniformly best quadratic estimators. Under the added assumption of normality, Graybill and Wortham (1956) showed these estimators are UMVU. A discussion of these results is given by Hocking (1985). Even in this ideal situation, the estimates are often unacceptable in the sense of violating the implicit assumption of nonnegativity. Several authors have proposed alternatives which guarantee nonnegative estimates, including Thompson and Moore (1963), Hartly and Rao (1967), Rao and Chaubey (1978) and Hartung (1981). Searle (1971ab) discusses various alternatives in some detail. Examples show spurious data can lead to negative estimates and Leone, et al (1968) have shown that negative estimates have non-trivial probability of occuring. The fact that spurious data can lead to negative estimates suggests that even positive estimates should be questioned and stresses the need for good diagnostic methods.

In the case of unbalanced data, there is a sharp discontinuity in theory. Except for special cases, minimal sets of sufficient statistics are not known, and, even in those special cases, they are not complete. Many estimators have been proposed and they fall generally into two categories. In one category are estimators based on quadratic forms, usually obtained from the mean squares of an AOV table. MINQUE and

related methods are included in this category. There is no basis to support the superiority of any of these approaches. Iterative methods fall into a second category and include maximum likelihood and REML. Other than large sample properties, little is known of the properties of these estimators. In addition, the iterative computations often encounter convergence difficulties.

The situation regarding the estimation of fixed effects parameters (means) is similar. With balanced data, the estimates are not affected by the presense of a non-scalar covariance matrix and they are UMVU estimators. With unbalanced data, maximum likelihood leads to weighted least squares estimators which depend on the unknown variance components. The properties of fixed effects estimators computed using estimated variance components are unknown.

The present paper discusses two contributions to the study of mixed models. First is the development of a new class of unbiased estimators for the case of unbalanced data which have simple, closed-form expressions. These expressions allow easy computation of variances which, when compared to minimum variance bounds, show the estimators to be highly efficient.

The second contribution discussed is the development of diagnostic methodology, based on the estimator, for assessing the effect of the data on the estimates. The source of negative estimates of variance components is often revealed by this methodology, as well as other sorts of instability and problems with the model or the design.

An overview of the methodology and its growing literature will be given. Applications of the ideas developed will be discussed in the context of several industrial problems for illustrative purposes. It is to be stressed that the methodology applies to random and mixed models, whether factorial or partially nested and whether balanced or unbalanced. Indeed, completely nested designs have been succesfully analyzed by this methodology by Hocking and M. S. Von Tress, but will not be discussed here. Also not discussed here is the distribution theory developed by Green and J. Grynovicki.

The problem of estimating variance components is shown to be equivalent to the problem of estimating the covariances between appropriate related observations. A covariance is naturally estimated by the corresponding sample covariance. In fact, almost every covarariance, θ_t, of the relevant sort can be estimated in an unbiased and efficient manner by a simple average of sample covariances, all having the same expectation and all simply related to θ_t, or else, by simple linear functions of such averages. In balanced cases, these sample covariances have the same distribution. In any case, they provide diagnostic power for examining the quality of the estimate of θ_t. The diagnostics are directly in terms of the effect influential factors have on parameter estimates of interest. Thus, only features of the data impacting on variance component estimates are highlighted. For small problems, these diagnostics are conveniently displayed in tables, as shown below. For larger problems, the diagnostics can be displayed in simple plots, as indicated below and described by Green (1987). For very large problems, reduction formulae, given by Green (1988) are available to reduce the demands of these displays to managable

levels. These are also discussed below. Since there are, in fact, many ways to generate meaningful diagnostics, these same formulae allow one to change from one representation to another, and even to increase the number of diagnostic elements.

2. THREE- AND FOUR-FACTOR MODELS

To motivate the procedure and introduce some general notation, consider a model with Factors 1, 2 and 3 (or 1, 2, 3 and 4) with Factor i having a_i levels. Let $r_i = a_i - 1$, $a_{12} = a_1 a_2$, $r_{12} = r_1 r_2$, etc. Let $r_0 = a_0 = 1$. Suppose there are $n_{ijk} \neq 0$ (or in the four-factor case, $n_{ijkt} \neq 0$) observations in the indicated cell. The empty cell problem will be reported on at a later date, although a brief discussion is given by Hocking (1987). Five model will be described to introduce the AVE-estimator and the diagnostic procedure. Two parameterizations are given. One is standard. The other is equivalent, but suggests both the diagnostic philosophy and the AVE-estimator, as well as an alternative statistical model which is more general than the usual model and has intuitive appeal.

2.1 Five Designs

To introduce the two parameterizations, consider the following three- and four-factor designs.

Design 1. Factors 1, 2 and 3 are crossed, 2 and 3 are fixed and 1 is random.

Design 2. Factors 1 and 3 are fixed and crossed, Factor 2 is random and nested in 1.

Design 3 is the same as Design 2, except Factor 1 is random.

Design 4 is the same as Design 2, except all factors are random.

Design 5. Factors 1, 3 and 4 are crossed, 2 is nested in 1, 1 and 2 are random and 3 and 4 are fixed.

2.2 Statistical Models for the Five Designs

In the case of design 1, a standard model is

(2.1) $y(ijks) = M(jk) + A(i) + AB(ij) + AC(ik) + ABC(ijk) + E(ijks)$,

where $M(jk)$ is the population mean of responses at levels jk of factors 23 and the others are independent 0-mean normal random variables with variances ϕ_1, ϕ_{12}, ϕ_{13}, ϕ_{123} and ϕ_0, respectively, and $y(ijks)$ is the s-th response at levels i, j, k of factors 1, 2, 3, respectively. It is useful to compute Θ_t, the covariance of distinct observations at the same level of factors indexed in t and at different levels of all other factors. Also, Θ will denote the total variance in the response. Thus, $\Theta = \phi_0 + \Theta_{123}$ in the three-factor case. The covariance structure in design 1 is given by:

(2.2) $Cov(y(ijks), y(i^*j^*k^*s^*)) =$

$$
\begin{cases}
0 & \text{if } i \neq i^* \\
\Theta_1 = \phi_1 & \text{if } i = i^*,\ j \neq j^*,\ k \neq k^* \\
\Theta_{12} = \phi_1 + \phi_{12} & \text{if } ij = i^*j^*,\ k \neq k^* \\
\Theta_{13} = \phi_1 + \phi_{13} & \text{if } ik = i^*k^*,\ j \neq j^* \\
\Theta_{123} = \phi_1 + \phi_{12} + \phi_{13} + \phi_{123} & \text{if } ijk = i^*j^*k^*,\ s \neq s^* \\
\Theta = \phi_0 + \Theta_{123} & \text{if } ijks = i^*j^*k^*s^*.
\end{cases}
$$

It should be observed that the parameterization given, in particular, the independence assumed of the "random effects", does not restrict the model. Rather, it indicates which of several equivalent

parameterizations is used. The covariance structures for the other designs follow.

Design 2.

(2.3) $y(ijks) = M(ik) + AB(ij) + ABC(ijk) + ABC(ijk)$,

where M(ik) is the population mean of levels ik of Factors 1 and 3, respectively, and the other terms are 0-mean normals with variances ϕ_{12}, ϕ_{123} and ϕ_0, respectively. The covariance structure is given in (2.4).

(2.4)
$$\begin{aligned} \Theta_{12} &= \phi_{12} \\ \Theta_{123} &= \phi_{12} + \phi_{123} \\ \Theta_0 &= \phi_0 + \Theta_{123} \end{aligned}$$

Design 3.

(2.5) $y(ijks) = A(i) + AB(ij) + M(k) + AC(ik) + ABC(ijk) + E(ijks)$,

where M(k) is the population mean of Factor 3, level k and the other terms are 0-mean normals with variances ϕ_1, ϕ_{12}, ϕ_{13}, ϕ_{123} and ϕ_0, respectively. The covariance structure is given in (2.6).

(2.6)
$$\begin{aligned} \Theta_1 &= \phi_1 \\ \Theta_{12} &= \phi_2 + \phi_{12} \\ \Theta_{13} &= \phi_1 + \phi_{13} \\ \Theta_{123} &= \phi_1 + \phi_{12} + \phi_{13} + \phi_{123} \\ \Theta_0 &= \phi_0 + \Theta_{123} \end{aligned}$$

Design 4.

(2.7) $y(ijks) = M + A(i) + AB(ij) + C(k) + AC(ik) + ABC(ijk) + E(ijks)$,

where M is the mean and the other terms are 0-mean normals with variances ϕ_1, ϕ_{12}, ϕ_3, ϕ_{13}, ϕ_{123}, ϕ_0. The covariance structure is in (2.8).

(2.8)
$$\Theta_1 = \phi_1$$

$$\Theta_{12} = \phi_1 + \phi_{12}$$
$$\Theta_{3} = \phi_3$$
$$\Theta_{13} = \phi_1 + \phi_3 + \phi_{13}$$
$$\Theta_{123} = \phi_1 + \phi_{12} + \phi_3 + \phi_{13} + \phi_{123}$$
$$\Theta_{0} = \phi_0 + \phi_{123}$$

Design 5.

(2.9) $$y(ijkts) = M(kt) + A(i) + AB(ij) + AC(ik) + ABC(ijk) + AD(it) + ABD(ijt) + ACD(ikt) + ABCD(ijkt) + E(ijkts),$$

where M(kt) is the population mean of responses at levels k and t of factors 3 and 4, respectively, and the other terms are independent, 0-mean normals with variances ϕ_1, ϕ_{12}, ϕ_{13}, ϕ_{123}, ϕ_{14}, ϕ_{124}, ϕ_{134}, ϕ_{1234} and ϕ_0, respectively. The covariances are given by (2.6), excluding 0, and by

(2.10)
$$\Theta_{134} = \phi_1 + \phi_{13} + \phi_{14} + \phi_{134}$$
$$\Theta_{1234} = \phi_1 + \phi_{12} + \phi_{13} + \phi_{123} + \phi_{14} + \phi_{124} + \phi_{134} + \phi_{1234}$$
$$\Theta_{0} = \phi_0 + \Theta_{1234}$$

with Θ_{14} and Θ_{124} analogous to Θ_{13} and Θ_{123}. It is evident that estimation of the Θ_t is equivalent to estimation of the ϕ_t. There are two advantages to the Θ_t parameterization. First, these covariances are rather naturally estimated by corresponding sample covariances. This estimation idea is the basis of AVE-estimator introduced (for unbalanced designs) in Hocking, Bremer and Green (1987), hereafter called (HBG). It is equivalent, in the balanced case, to the usual ANOVA estimator (HBG), Green (1985, 1988) and offers an efficient Hocking (1987), (HBG) alternative in the unbalanced case. A second advantage is the possibility of a more general formulation of the model in terms of the mean and covariance structure of the response vector, y. For example, in design 1, the model can be specified by writing

E[y(ijks)] = M(jk) and COV (y), as given by (2.2).
The only restriction on the covariance structure is that the covariance matrix be positive definite. This is true if all the ϕ_t are positive, but also under more general conditions which permit individual "variance" components to be negative. Explicit requirements for positive definitness are given in Hocking (1985). Since physically, a negative covariance is possible (See Green(1988), for an industrial setting in which a negative covariance is quite sensible), this more general formulation has some appeal. It also provides an explanation for the negative variance component estimates which frequently occur. The validity of the AVE-estimator or the diagnostic procedure does hinge on acceptance of this alternative model.

2.3 Estimation of Variance Components Arising from the Five Designs

It is natural to estimate the covariances θ_t by corresponding sample covariances. This is the basis of the diagnostic procedure. In the balanced case, the estimates found are the usual estimates obtained from an AOV table (Henderson's type H3 or SAS type 2).

Some simple notation is introduced to facilitate the procedure. The general form is given in Green (1987, 1988), (HBG) appropriate for any design. For the present, forms needed for three or four factors are given. These contain all the basic forms required in general. They are not tied to any particular design.

To estimate the covariance, θ_1, between observations at the same level of Factor 1 but different levels of Factors 2 and 3, one of the

following three forms is used.

$$(2.11)\quad C(1/23) = (a_{23}\, r_{23})^{-1} \sum C(1/jkj^*k^*).$$

$$(2.12)\quad C(1/2{:}3) = (a_{23}\, r_{2}\, a_{3})^{-1} \sum C(1/jkj^*k^*).$$

$$(2.13)\quad C(1/3) = (a_{3}\, r_{3})^{-1} \sum C(1/kk^*).$$

In (2.11), the sum is over all $a_2 r_2$ pairs of distinct levels $j \neq j^*$ of Factor 2 and all $a_3 r_3$ pairs of distinct levels $k \neq k^*$ of Factor 3. In (2.12), the sum is over the $a_2 r_2$ pairs of distinct levels of Factor 2 and all $a_3 \times a_3$ pairs of levels of Factor 3, whether or not distinct. In (2.13) the sum is over the $a_3 r_3$ pairs of distinct levels of Factor 3. In (2.11) and (2.12),

$$(2.14)\quad C(1/jkj^*k^*) = r_1^{-1} \sum_i \left(\bar{y}(ijk.) - \bar{y}(.jk.) \right)\left(\bar{y}(ij^*k^*.) - \bar{y}(.j^*k^*.) \right),$$

where $\bar{y}(ijk.)$ is a cell mean and $\bar{y}(.jk.)$ is an (unweighted) mean of cell means. (2.14) is a sample covariance of cell means at the same level of Factor 1 and at indicated levels of Factors 2 and 3. (2.13) is the average of forms of the sort (2.15), which is a sample covariance of the average responses of Factor 1 at indicated levels of Factor 3.

$$(2.15)\quad C(1/kk^*) = a_2^{-2} \sum_{jj^*} C(1/jkj^*k^*)$$
$$= r_1^{-1} \sum_i \left(\bar{y}(i.k.) - \bar{y}(..k.) \right)\left(\bar{y}(i.k^*.) - \bar{y}(..k^*.) \right)$$

Justification for using unweighted means of the cell means in the unbalanced case is discussed in (HBG) and is as follows. Begin with the balanced case, where the forms are clearly reasonable. (HBG) shows that in the unbalanced case, if one uses these forms for all possible balanced submodels of minimum cell frequency and averages these estimators over all such submodels, the resulting average is the AVE-estimator as described here. Which of the forms to use in a problem

is determined by the nesting and fixed factors in the design and is explored below.

To estimate the covariance, θ_{12}, between observations at the same level of Factors 1 and 2 but different levels of Factor 3, (2.14) or one of the following two forms is used (in a three-factor model).

$$(2.16)\quad C(12/3) = (a_3 r_3)^{-1} \sum C(12/kk^*),$$

$$(2.17)\quad C(1,2/3) = (a_{13} r_3)^{-1} \sum C(i,2/kk^*),$$

where the first sum is over all $a_3 r_3$ pairs of distinct levels $k \neq k^*$ of Factor 3 and the second sum is also over these and over all a_1 distinct levels of Factor 1. Here,

$$(2.18)\quad C(12/kk^*) = r_{12}^{-1} \sum_{ij} (\bar{y}(ijk.)-\bar{y}(..k.))(\bar{y}(ijk^*.)-\bar{y}(..k^*.))),$$

$$(2.19)\quad C(i,2/kk^*) = r_2^{-1} \sum_{j} (\bar{y}(ijk.)-\bar{y}(i.k.))(\bar{y}(ijk^*.)-\bar{y}(i.k^*.))).$$

In all forms, by permutation of the indicies, one obtains analogous forms appropriate for estimating the other covariances. Now consider the five designs stated above.

Design 1.

$$\tilde{\theta}_1\text{-AVE} = C(1/23)$$

$$(2.20)\quad \tilde{\theta}_{12}\text{-AVE} = C(2,1/3)$$

$$\tilde{\theta}_{13}\text{-AVE} = C(3,1/2)$$

Design 2.

$$(2.21)\quad \tilde{\theta}_{12}\text{-AVE} = C(1,2/3)$$

Design 3.

$$\tilde{\theta}_1\text{-AVE} = C(1/3) - a_2^{-1} C(1,2/3)$$

(2.22)

$$\tilde{\theta}_{12}\text{-AVE} = C(1/3) + r_2 a_2^{-1} C(1,2/3).$$

Design 4.

$\tilde{\theta}_1$ -AVE and $\tilde{\theta}_{12}$ -AVE are as in Design 3.

(2.23) $\tilde{\theta}_3$- AVE = C(3/1:2)

$$\tilde{\theta}_{13}\text{-AVE} = C(1,3/2) + \tilde{\theta}_1\text{-AVE}.$$

Design 5.

(2.24) $\tilde{\theta}_1$ -AVE = $C(1/34) - a_2^{-1} * C(1,2/34)$

$\tilde{\theta}_{12}$ -AVE = $C(1/34) + r_2 a_2^{-1} * C(1,2/34)$

$\tilde{\theta}_{13}$ -AVE = $C(3,1/4) - a_2^{-1} * C(13,2/4)$

$\tilde{\theta}_{123}$ -AVE = $C(3,1/4) + r_2 a_2^{-1} * C(13,2/4)$

The estimators for the 14-and 124-interactions are obtained from those for 13 and 123 by interchange of indicies.

In all cases, the highest order term (θ_{123} or θ_{1234}) suggests no sample covariance estimator, since, if the model is correct, the order of observations within a cell is arbitrary. Also, some terms of highest order in the non-nested factors are not well-represented by sample covariances of the obvious type. However, an AVE-type estimator can be be based on deletion methods. Such are discussed in (HBG).

2.4 Estimation of Fixed Effects

Similar unweighted means are used to estimate the fixed effects.

It will be noted that the estimators of the fixed effects in a balanced design are linear combinations of the cell means. The idea of averaging over all possible designs of minimum cell size (provided that size is not zero) leads to the same linear combination, except that with unbalanced data, the cell means are based on different numbers of observations. The result is to replace an expression such as

$$(2.25)\quad M(ij)^{\sim} = \sum_{ks} (a_3 n)^{-1} y(ijks)$$

in the balanced case by

$$(2.26)\quad M(ij)^{\sim}\text{-AVE} = a_3^{-1} \sum_k (n_{ijk})^{-1} \sum_s y(ijks).$$

(HBG) contains a discussion of fixed effects estimation in unbalanced factorial models. Hocking (1987) continues this discussion, with reference to partially nested models. Further joint work on this latter topic is expected to appear soon.

2.5 Display and Use of Diagnostics

Now that the basic forms are evident, attention can turn to their use. Each term C(p,v/d) is an average of sample covariances, all of which have the same expectation. In design 1, the general representation theorem Green (1988) gives the forms (2.20). The AVE-estimator of $\Theta_1^{\sim}$ is C(1/23), which is the average of the $a_{23} r_{23}/2$ distinct sample covariances C(1/jkj*k*), for $j \neq j^*$, $k \neq k^*$. Each of these covarianaces is an unbiased estimate of $\Theta_1^{\sim}$. They can be displayed in a table, such as Table 1, which shows $a_2 = 2$ and $a_3 = 4$. In this illustration, one 4-by-4 table gives all the diagnostics. The off-diagonal elements are the sample covariances. Since this table is not symmetric, all off-diagonal

elements are printed. The diagonal elements are not true variances, since j = 1 and j* = 2 there, while k = k*. If two tables were given, one could compute and report the following variances.

$$(2.27) \qquad C(1/jkjk) = r_2^{-1} \sum_j (\bar{y}(ijk.) - \bar{y}(.jk.))^2 .$$

Under the usual assumptions for this design, all diagonal elements have the same expectation, as do all off-diagonal elements. The table is examined for outliers and patterns. Green (1988) gives moments of these diagnostic elements. In this example, the elements C(1/jkj*k*) for jk, j*k* = 12, 13 and 13, 22 stand out as much larger than the other entries. Also, the diagonal entries for k = 2 and k = 3 are much larger than the other diagonal entries. This suggests further examination of the two combinations indicated. In a paper presented at the Gordon Research Conference, August, 1987, and being prepared by the present authors for publication, this table was part of an analysis which detected a process shift in data from an actual chemical production process. This point will be elaborated on below. One use of such tables is the detection of problems in the underlying assumptions made about the model. One conclusion drawn for the chemical data is that a violation of this sort occurs. A physical consequence is the need to redesign the production line to make a uniform product.

A second application of these diagnostic tables is the detection of spurious data. The second point is illustrated in the context of a wool fiber example discussed in Green (1987). The design is Design 2, with $a_1 = 2$, $a_2 = 5$, $a_3 = 23$. The estimate, C(1,2/3), of $\tilde{\Theta}_{12}$ is the average of the sample covariances C(i,2/kk*), i = 1,2 and $k \neq k^* = 1,\ldots,23$. A tabular display of these diagnostics would require two 23-by-23 tables, an unpleasant prospect. In the above cited article, these

diagnostics are displayed in simple plots. For each value of k, the value C(i,2/kk*) is plotted against k*, using as plotting symbol the value of i. Figure 1 is the result for k = 5 and k = 6. Two features stand out in this plot. First, the i = 2 values are almost all higher than the corresponding i = 1 values. Second, the value for i = 2 and k* = 11 are dramatically higher than than all other points, for both k = 5 and k = 6. A logical followup step is to plot $\bar{y}$(ijk.) vs $\bar{y}$(ijk*.) for i = 2, k = 6, k* = 11 and all j. This is done in Figure 2. The plotting symbol is A for j = 1, B for j = 2, etc when i = 2, and 1 for j = 1, 2 for j = 2, etc when i = 1. If a sample covariance appearing in the table is a stable estimate of Θ_{12}, one would expect a clear linear trend in the plot of cell means. At i = 2, there is evident a serious problem due to the effect of j = 1 (the point A). Serious reservations about the sampling methodology are raised in the article by this (and similar) points.

To return to the chemical data, a followup plot of cell means at jk = 13 vs jk = 22 is shown in Figure 3. Here there is a strong linear trend to the points, unlike the wool example. A possibly spurious point is seen, but deletion of this point has little effect on the Θ_1 estimate. The plotting symbol used indicates in which level of factor 1 the data point falls. The symbol 0 is for i = 1-10, 1 is for i = 11-20, etc. The plot suggests the higher levels of i give higher points. A subsequent plot (Figure 4) of cell means $\bar{y}$(ijk.) vs i, for jk = 12 (and jk = 13, 22, 23 are similar) shows a pronounced shift at around i = 30. (a_1 = 60 in this problem.) Process engineers verified a change in raw material at this point.

2.6 Reductions in Size of Displays

If neither tabular nor graphical display seems feasible, Green (1988) offers algebraic reduction formulae and partial summing methods, which, together with the general moment formulae developed there, allow smaller tables to be constructed which retain most of the diagnostic power suggested by these examples. He describes a six factor design which would require the display of 15,680 sample covariances. This seems an unreasonable demand. The reduction formulae cut the required display to 840 sample covariances, a reduction of 94 %. Further reductions are possible through partial summing of the diagnostic forms, as described in the context of a glass manufacturing example.

Consider now design 5, with diagnostic forms given by (2.24). Green (1988) considers a glass manufacturing example with $a_1 = a_3 = 5$, $a_2 = 2$, $a_4 = 3$. These forms require displaying 630 diagnostic elements. After applying the reduction formulae, a display of C(13,2/4) is still required. Conceptually, the terms C(ik,2/tt*) are displayed in table form. Perhaps, for each value of i and k, an a_4-by-a_4 table is constructed, the off-diagonal terms of which are the sample covariances. The below-diagonal terms need not be displayed, since the table is symmetric. Diagonal entries are sample variances, which also carry diagnostic information. In the example, this requires 25 3-by-3 tables, a rather onerous requirement. The graphical displays discussed above can be used if the number of terms is moderate. Even these displays may be problematic for larger values of a_1, a_3 and a_4. A simple remedy is to work with "partial sums" described below. In the glass example, the forms (2.24) can be replaced by:

(2.28)

C(1,2/4), with 15 elements

C(1/4), with 3 elements

C(13,2/4), with 30 elements through

$$a_1^{-1} \sum_i C(ik,2/tt^*) \quad \text{and} \quad a_3^{-1} \sum_k C(ik,2/tt^*)$$

(2.28a) (2.28b)

C(14,2/3), with 80 elements through

$$a_1^{-1} \sum_i C(it,2/kk^*) \quad \text{and} \quad a_4^{-1} \sum_t C(it,2/kk^*)$$

C(3,1/4), with 15 elements

C(4,1/3), with 9 elements.

This gives a total of 152 diagnostic elements, a reduction of 75 %. As shown by Green (1988), the remaining elements have essentially the same diagnostic power as a full analysis. Further reduction is possible in the last two terms. In this example, there are so few diagnostics in in these two that further reduction makes little sense.

The analysis now is in four parts. (1) Outlier analysis associated with each table finds those estimates more than 2σ away from the mean for that table. (2) In the case of tables for the partial sums, if, say, for some i, one of the off-diagonal terms in (2.28a) stands out, then a table of C(ik,2/tt*) for just that i is constructed, or else a univariate analysis of the estimates C(ik,2/tt*) is done (either using stem-and-leaf plots or a printout of values outside a 2- or 3-σ confidence band). (3) Next, a "pattern analysis" of the tables may bring out special patterns. There should be no pattern to the tables if the statistical model assumptions are correct. (4) Next, the the data set is examined to seek statistical cause for what was seen

(1)-(3).

Each table shows appropriate, equal-expectation, sample covariances off the diagonal. Since these tables are symmetric, below diagonal terms are omitted. The diagonal terms in these tables are variances, and always have equal expectations under standard assumptions.

To continue with the illustration, Table 2 gives the diagnostics (2.28a). The entry for i = 5, tt* = 12 stands out a large. This can be judged by inspecting either the table or a stem-and-leaf plot, or with the aid of a 2σ confidence band centered at the average value of (2.28a). In this last regard, the following variance formula is helpful.

(2.29)

$$\text{VAR (Form 2.28a)} = (a_3 r_2)^{-1} [\, a_3 (\theta_{124} - \theta_{14})^2 + r_3 (\theta_{12} - \theta_1)^2 + (\theta - \theta_{134})^2 \,].$$

From this, the standard deviation is 311.0 and the average value is seen to be 329.5 Similar computations apply to the diagonals. A printout of the forms C(ik,2/tt*) for i = 5, tt* = 12 outside a 3σ confidence band shows k = 3 and k = 4 account for the initial large estimate. This in turn leads to an examination of the relevant data, where a large difference between the values for j = 1 and j = 2 is found at these locations. A complete discussion of this data is given in the cited article, but this should indicate how the "reduction" techniques work.

In connection with the above analysis, the first two momemts of the diagnostic forms involved in (2.28) are needed. General closed-form expressions for moments of the required type are given. These are functions of the θ_t and apply to balanced and unbalanced cases.

2.7 Repeated Measures Experimrnts

Grynovicki and Green (1988) contains a discussion of this methodology to repeated measures experiments. In the example described there, the diagnostics lead to the discovery of two populations of subjects not properly taken into account in the study and which raise serious questions about the validity of conclusions to be drawn. The existence of these two populations had not been previously suspected. Applications to other repeated measures experiments, such as medical experiments, are readily apparent.

2.8 Computations

The computations involved in constructing the tables or plots presented above are minimal. Standard statistical computer packages will do all calculations required, though some manipulation may be required to print the diagnostic tables in a useful format. For example, SAS PROC CORR, with the COV option will compute sample covariances and even display them, often in appropriate form. The plots require additional data manipulation, but again standard packages have the requisite capability. All computations discussed here were done using SAS.

The reduction and partial summing ideas discussed make this methodology applicable to designs of all sizes. Since the methodology also applies regardless of the degree of imbalance and to a large class of mixed models, it can be seen to be useful in a wide variety of problems.

2.9 Efficiency of the AVE-Estimator

(HBG) and, more definitively, Hocking (1987) contain discussions the efficiency of the AVE-estimator. This is done by comparing the small sample variances of these estimators with lower bounds for this variance, as given by Bhattacharya (1946) in an improvement of the usual Cramer-Rao lower bounds. Closed-form expressions for these bounds are not known, but they can be computed numerically for specific designs. Such computation is reported in the cited articles for a variety of cell frequency patterns and parameter values. Among the conclusions reported there are the following.

1. The AVE-estimators of both variance components and fixed effects are very efficient.
2. The efficiencies are monotonically increasing in all parameters.
3. The efficiencies depend on all parameters but the variances do not.
4. When compared to Yates' method (or the method of weighted square of means or SAS type 3) or Henderson's method (or the method of fitting constants or SAS type 2), there is little reason to distinguish among these estimators on the grounds of efficiency, although the AVE-estimator is generally superior except for small parameter values.

3. OTHER LITERATURE

The first article on the general diagnostic philosophy described was Hocking (1983) which applied these ideas to balanced randomized block designs. Alternative models, such as discussed above, which allow for negative estimates of variance components, were discussed by

Smith and Murray (1984) for certain two-factor models, but no diagnostics were described there. The first major development of diagnostics was given by Green (1985), a dissertation written under the direction of Hocking. Results based in part on this were reported by Hocking (1985) and Hocking and Pendelton (1985). It deals with balanced, random models only, but, with minor changes, applies to mixed models. Matrix expressions for various diagnostic forms and moments are given which simplify computations by hand or computer for balanced designs. Since most diagnostic forms in the unbalanced case are unweighted linear functions of the cell means, many results from the balanced case apply with little or no change to the unbalanced case. Hocking and Bremer were the first to notice the unbalanced extension. Some results from this source will appear in a more available format in the near future. Results from (HBG) are discussed in (HBGb), although in the conference proceedings, an administrative error omitted one author's name.

4. CONCLUSIONS

A diagnostic procedure has been shown to be both intuitively simple and effective in judging the quality of variance component estimates. It applies to both small and large problems. All calculations, displays and plots can be (and were) done by standard statistical computing packages. The diagnostics are themselves estimates of the components in question, and, as such, indicate in a straight forward manner, what impact various features of the data have on the overall estimates. Only features of the data affecting the parameter estimates are flagged. The methodology applies to both balanced and unbalanced designs with no missing cells. A sound theoretical basis exists for the procedure. In

the balanced case, the overall estimator based on the diagnostics is a standard one obtained from equating mean squares to expected mean squares, whereas in the unbalanced case, the estimator is new and compares favorably with standard estimators in terms of efficiency. In addition, in the unbalanced case, the estimator is in closed form, which simplifies both computation and theoretical inquiry. Also of importance is the fact the method applies in any random or mixed model to all components of variance other than the highest order in the non-nested factors, and even to some of these, without modification, as well to fixed factors. With some modification, these estimates apply to these highest order terms as well.

The diagnostic methodology brings out many noteworthy features of the data directly in terms of their effect on parameters of interest. Even for large data sets, the tabular and computational requirements are modest. The reduction formulae and univariate confidence interval approach reduce the need for tabular displays to a reasonable level. Unbalanced models are handled in the same way as balanced models, and with little added trouble. The methodology is sufficiently flexable to allow the user to tailor some computations to suit the needs of a particular problem, yet sufficiently standardized to be easily learned or programmed.

TABLE 1. Diagnostics C(i/jkj*k*) for θ_1

Chemistry Data

j = 1

k

		1	2	3	4
j* = 2 k*	1	6.8	16.7	14.6	9.7
	2	11.7	31.5	33.2	12.2
	3	10.5	31.1	27.1	17.3
	4	8.1	16.9	22.6	11.3

STEM & LEAF

```
30 | 23
20h|
20 | 3
10h| 5777
10 | 0122
 0 | 8
```

TABLE 2. Diagnostics $a_3^{-1} \sum_k C(ik,2/tt^*)$

Glass Data

STEM	LEAF
10	0
9	
8	2
7	
6	0
5	55
4	0
3	01
2	4
1	37
0	
-0	0348

	t* 1	2	3	
1	131.2	-43.5	-28.9	
t 2		156.0	- 1.2	i = 1
3			44.5	

	t* 1	2	3	
1	1582.2	821.6	546.8	
t 2		517.7	304.9	i = 2
3			392.5	

	t* 1	2	3	
1	89.2	242.2	170.7	
t 2		761.4	599.0	i = 3
3			593.3	

	t* 1	2	3	
1	241.5	133.6	-82.0	
t 2		586.6	396.4	i = 4
3			594.0	

	t* 1	2	3	
1	1631.8	1023.0	554.0	
t 2		893.2	306.4	i = 5
3			392.1	

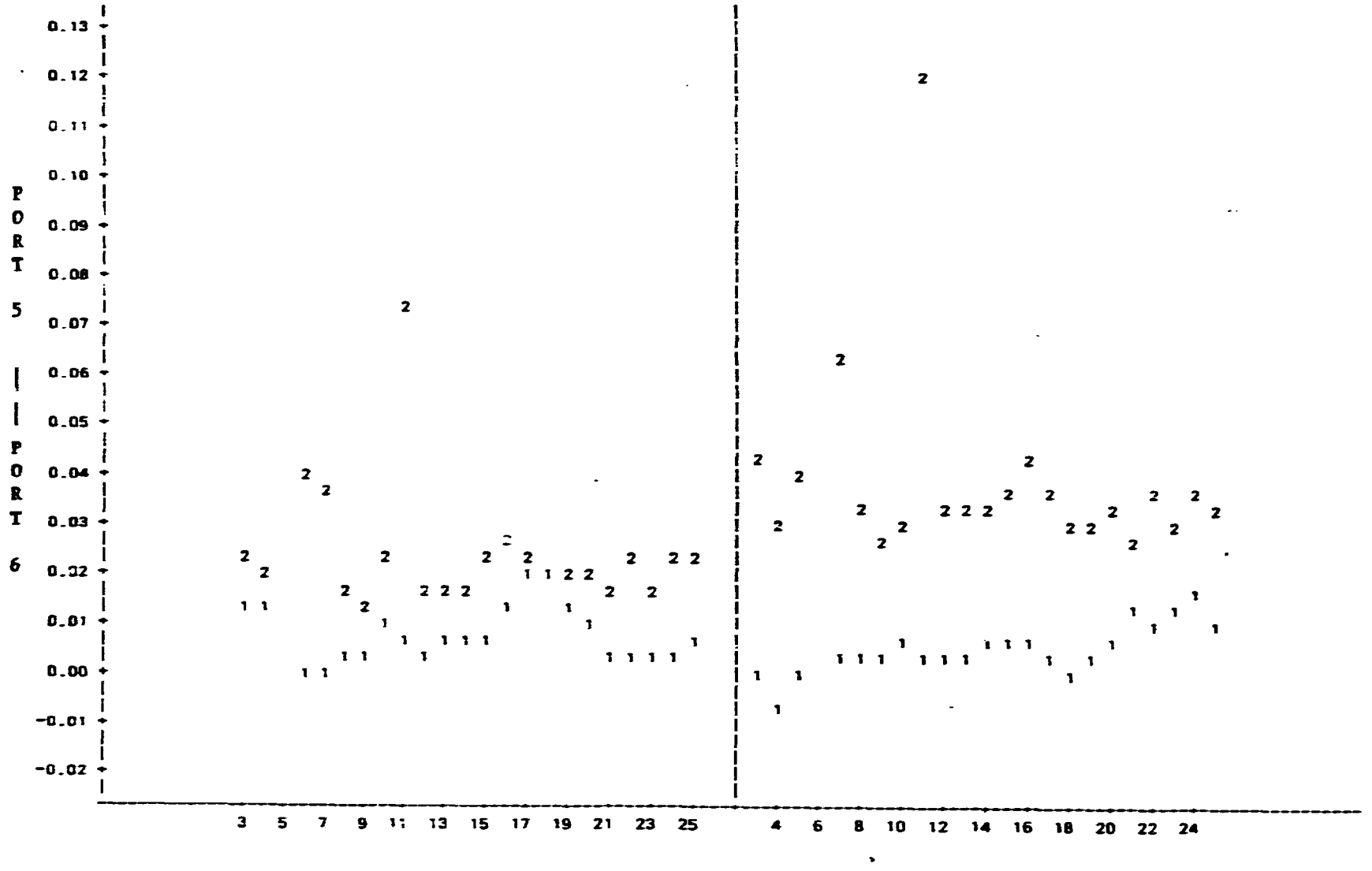


FIGURE 1. Wool Data Diagnostics C(i,2/kk*) for k = 5 and 6. Plotting symbol is level of i. Horizontal axis is value of k*.

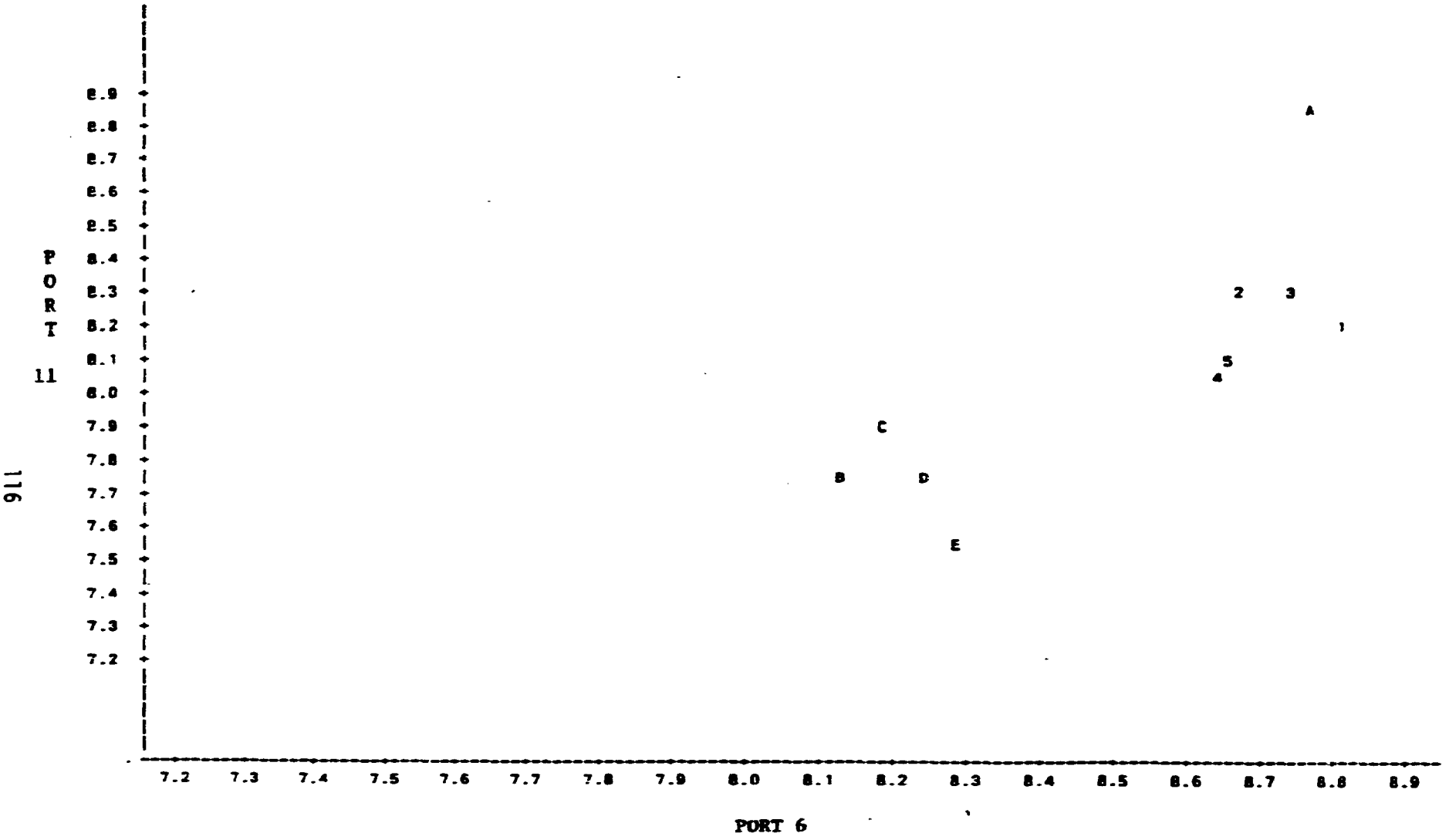


FIGURE 2. Wool Data. Cell means of k = 11 vs k = 6. The values for i = 2 denoted by letters, those for i = 1 by numbers.

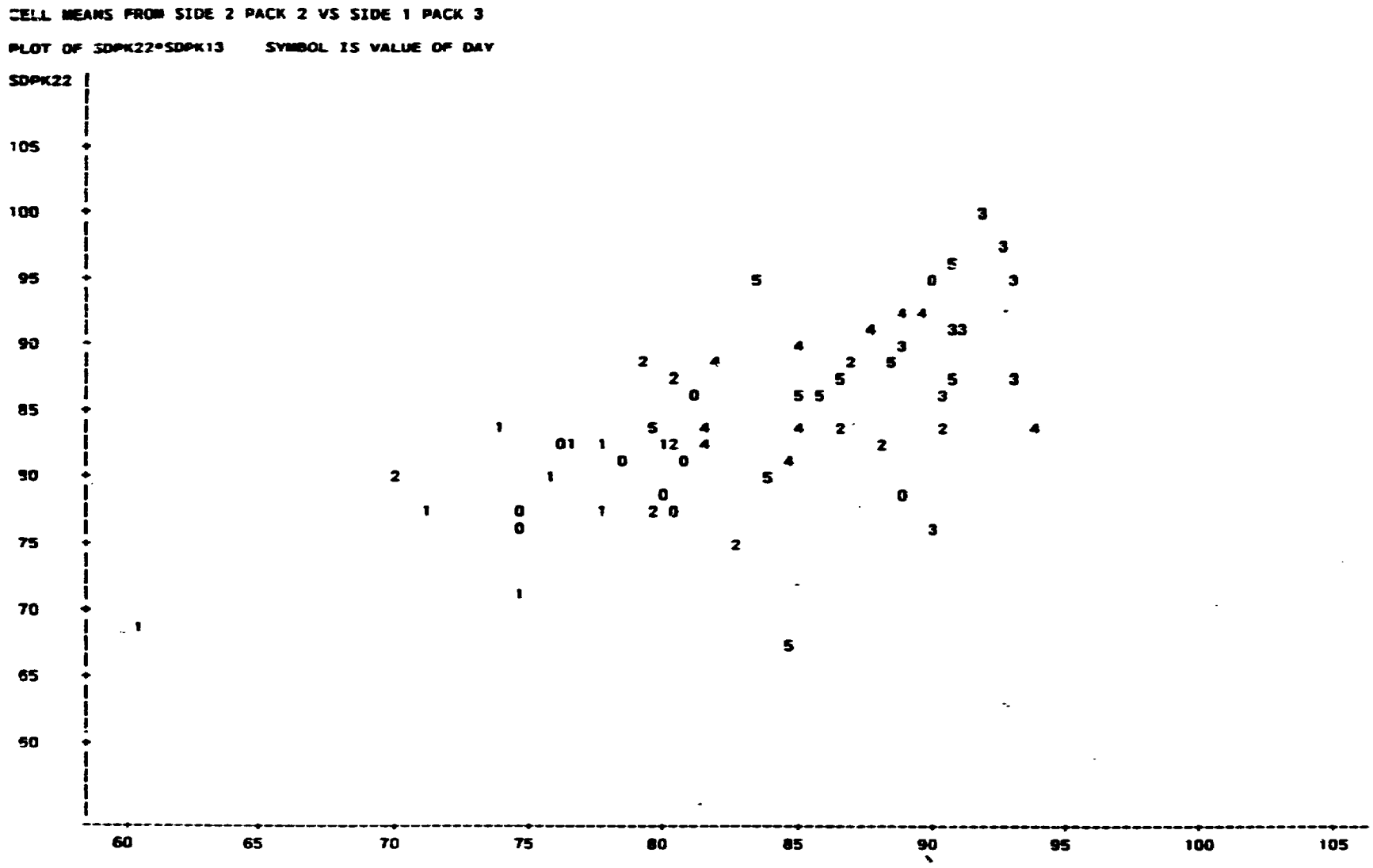


FIGURE 3. Chemistry Data. Cell means of jk = 22 vs jk = 13. Plotting symbol: 0 is for i = 1-10, 1 is for i = 11-20, etc. Factor 1 (i) has 60 levels.

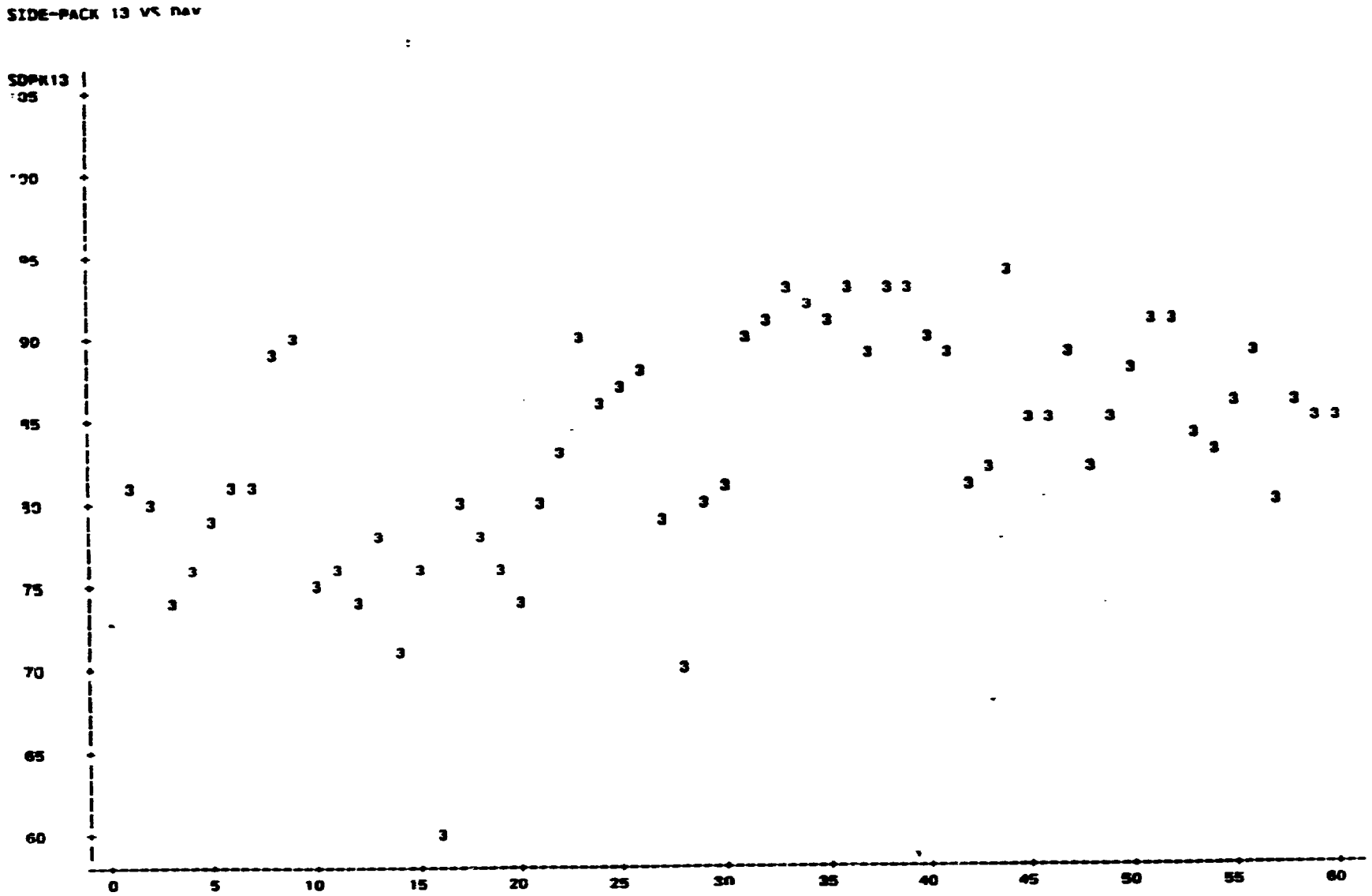


FIGURE 4. Chemistry Data. Cell means of jk = 13 vs i.

REFERENCES

Bhattacharyya, A. (1946), "On some Analogs of the Amount of Information and Their Use in Statistical Estimation", Sankhya 8, 1-14,201-208, 277-280.

Graybill,F.A. and Hultquist,R.A. (1961), "Theorems Concerning Eisenhart's Model II", Annals of Mathematical Statistics 32, 261-269.

Graybill, F.A. and Wortham, A.W. (1956), "A note on uniformly best Unbiased Estimators for Variance Components", JASA 51, 266-268.

Green, J. W. (1985), Variance Components: Estimates and Diagnostics, Dissertation, Texas A & M University

Green, J.W. (1987), "Diagnostic Procedure for Variance Components for Both Large and Small Designs", submitted to Technometrics.

Green, J.W. (1988), "Diagnostics for Variance Components Suitable for Designs of all sizes", submitted to Technometrics.

Grynovicki, J.O. and Green, J.W. (1988), "Estimation of Variance Components and Model-Based Diagnostics in a Repeated Measures Design", Proceedings of the Thirty-Third Conference on the Design of Experiments.

Hartley, H.O. and Rao, J.N.K. (1967), "Maximum Likelihood Estimation for the Mixed Analysis of Variance Model", Biometrics 54, 93-108.

Hartung, Joachim (1981), "Nonnegative Minimum Biased Invariant Estimation in Variance Component Models", The Annals of Statistics 9, 278-292.

Harville, D.A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems", JASA 72, 320-340.

Hocking, R.R. (1983), "A Diagnostic Tool for Mixed Models with Applications to Negative Estimates of Variance Components", Proceedings of the SAS Users Group, New Orleans, LA., 711-716.

Hocking, R.R. (1984), "Diagnostic Methods in Variance Component Estimation", Proceedings of the XII-th International Biometrics Conference, Tokyo, Japan, 49-58.

Hocking, R.R. (1985), The Analysis of Linear Models, Monterrey,CA, Brooks-Cole

Hocking, R.R. (1987), "A Cell Means Analysis of Mixed Linear Models", submitted.

Hocking, R.R., Bremer, R.H. and Green,J.W. (1987), "Estimation of Fixed Effects and Variance Components in Mixed Factorial Models Including Model-Based Diagnostics", submitted to Technometrics.

Hocking, R.R, and Green,J.W. (1985), "New Expressions for Variance Component Estimators with Diagnostic and Modeling Implications", unpublished manuscript.

Khuri, A.I. and Sahai, H. (1984), "Variance Components Analysis: A Selective Literature Survey", Technical Report No. 226, Department of Statistics, University of Florida, Gainesville, Florida.

Leone, F.C. and Nelson, L.S. (1965), "Sampling Distributions of Variance Components. I. Empirical Studies of Balanced Nested Designs", Technometrics 8, 457-468.

Leone, F.C., Nelson, L.S., Johnson, N.L. and Eisenstat, S. (1968), "Sampling Distributions of Variance Components, II. Empirical Studies of Unbalanced Nested Designs", Technometrics 10, 719-738.

Nelder, J.A. (1954), "The interpretation of Negative Components of

Variance", Biometrika 41, 544-548.

Rao, P.S.R.S. and Chaubey, Y.P. (1978), "Three Modifications of the Principle of the Minque", Communications in Statistics 47, 767-778.

Sahai, H. (1979), "A Bibliography on Variance Components", International Statistical Reviews 47, 177-222.

Sahai, H. and Khuri, A.I. (1984), "A Second Bibliography on Variance Components", Technical Report No. 217, Department of Statistics, University of Florida, Gainesville, Florida.

Searle, S.R. (1971), Linear Models, New York, John Wiley & Sons

Searle, S.R. (1971), "Topics in Variance Component Estimation", Biometrics 27, 1-76.

Smith, D.W. and Murray, L.W. (1984), "An Alternative to Eisenhart's Model II and Mixed Model in the Case of Negative Variance Estimates", JASA 79, 145-151.

Snedecor, G.W. and Cochran, W.G. (1963), Statistical Methods, Ames, Iowa, Iowa State Press.

Thompson, W.A. Jr. and Moore, J.R. (1963), "Non-Negative Estimates of Variance Components", Technometrics 5, 441-449.

THEORY OF SEMIREGENERATIVE PHENOMENA

N.U. Prabhu
School of Operations Research and Industrial Engineering
and Mathematical Sciences Institute
Cornell University, Ithaca, NY 14853, U.S.A.

Abstract: We develop a theory of semiregenerative phenomena. These may be viewed as a family of linked regenerative phenomena, for which Kingman ([6],[7]) developed a theory within the framework of quasi–Markov chains. We use a different approach and explore the correspondence between semiregenerative sets and the range of a Markov subordinator with a unit drift (or a Markov renewal process in the discrete time case). We use techniques based on results from Markov renewal theory.

Keywords: Semiregenerative phenomena and sets, linked regenerative phenomena, quasi–Markov chains, standard phenomena, stable states, lifetime, Markov renewal processes, Markov additive processes.

Introduction and Summary. Let the set T be either $[0,\infty)$ or $\{0,1,2,...\}$, E a countable set and $(\Omega, \mathcal{F}, P)$ a probability space.

Definition 1. A semiregenerative phenomenon $Z = \{Z_{t\ell}, (t,\ell) \in T\times E\}$ on a probability space $(\Omega, \mathcal{F}, P)$ is a stochastic process taking values 0 or 1 and such that for $(t_r,\ell_r) \in T\times E$ $(r \geq 1)$, with $0 = t_0 \leq t_1 \leq ... \leq t_r$, $j \in E$ we have

$$P\{Z_{t_1\ell_1} = Z_{t_2\ell_2} = ... = Z_{t_r\ell_r} = 1 \,|\, Z_{0j} = 1\}$$
$$= \prod_{i=1}^{r} P\{Z_{t_i - t_{i-1},\ell_i} = 1 \,|\, Z_{0,\ell_{i-1}} = 1\} \quad (\ell_0 = j). \tag{1}$$

For each $\ell \in E$, denote $Z_\ell = \{Z_{t,\ell}, t \in T\}$. Since

$$P\{Z_{t_1\ell} = Z_{t_2\ell} = ... = Z_{t_r\ell} = 1 \,|\, Z_{0j} = 1\}$$
$$= P\{Z_{t_1\ell} = 1 \,|\, Z_{0j} = 1\} \prod_{i=2}^{r} P\{Z_{t_i - t_{i-1},\ell} = 1 \,|\, Z_{0\ell} = 1\}, \tag{2}$$

Z_ℓ is a (possibly delayed) regenerative phenomenon in the sense of Kingman [7] in the continuous time case $T = [0,\infty)$, and a recurrent event (phenomenon) in the sense of Feller [5] in the discrete time case $T = \{0,1,2,...\}$. The family $Z' = \{Z_\ell, \ell \in E\}$ is a family of linked regenerative phenomena, for which a theory was developed by Kingman [6] in the case of finite E; later he reformulated the results in terms of quasi–Markov chains (Kingman [7]). We explain this concept below.

Example 1. Let $J = \{J_t, t \in T\}$ be a time–homogeneous Markov chain on the state space E and denote

$$Z_{t\ell} = 1_{\{J_t = \ell\}} \text{ for } (t,\ell) \in T \times E. \tag{3}$$

The random variables $Z_{t\ell}$ satisfy the relationship (1), which is merely the Markov property. More generally, let C be a fixed subset of E and

$$Z_{t\ell} = 1_{\{J_t = \ell\}} \text{ for } (t,\ell) \in T \times C. \tag{4}$$

These random variables also satisfy (1) and thus $Z = \{Z_{t\ell}, (t,\ell) \in T \times C\}$ is a semiregenerative phenomenon. In particular, suppose that C is a finite subset of E and define

$$K_t = J_t \text{ if } J_t \in C, \text{ and } = 0 \text{ if } J_t \notin C. \tag{5}$$

Then $\{K_t, t \in T\}$ is a quasi–Markov chain on the state space $C \cup \{0\}$. □

While the quasi–Markov chain does provide a good example of a semiregenerative phenomenon (especially in the case of finite E), it does not reveal the full features of these phenomena; in particular, it does not establish their connection with Markov additive processes. Thus, let

$$\zeta = \{(t,\ell) \in T \times E: Z_{t\ell} = 1\}. \tag{6}$$

We shall call ζ the semiregenerative set associated with Z. The main theme of this paper is the correspondence between the set ζ and the range of a Markov renewal process (in the discrete time case) and of a Markov subordinator with a unit drift (in the continuous time case). Kingman ([7], p. 123) has remarked that associated with a quasi–Markov chain there is a process of type F studied by Neveu [9]. The Markov subordinator we construct for our purpose is indeed a process of type F, but we concentrate on properties of the range of this process. For a detailed description of Markov additive processes see Cinlar ([2],[3]).

To complete Definition 1 we specify the initial distribution $\{a_j, j \in E\}$, where

$$P\{Z_{0j} = 1\} = a_j \tag{7}$$

with $a_j \geq 0$, $\Sigma a_j = 1$. As in the case of regenerative phenomena, it can be proved that the relation (1) determines all finite dimensional distributions of Z and that Z is strongly regenerative (that is, (1) holds for stopping times). We shall write P_j and E_j for the probability and the expectation conditional on the event $\{Z_{0j} = 1\}$.

In the discrete time case we call Z a semirecurrent phenomenon and denote

$$u_{jk}(n) = P\{Z_{nk} = 1 \,|\, Z_{0j} = 1\} \tag{8}$$

where $u_{jk}(0) = \delta_{jk}$. In the continuous time case let

$$P_{jk}(t) = P\{Z_{tk} = i \mid Z_{0j} = 1\} \tag{9}$$

where $P_{jk}(0) = \delta_{jk}$. The phenomenon is standard if

$$P_{jk}(t) \to \delta_{jk} \text{ as } t \to 0+. \tag{10}$$

In this case it is known that the limit

$$\lim_{t \to 0+} \frac{1 - P_{jj}(t)}{t} \quad (j \in E) \tag{11}$$

is known to exist (possibly infinite); if this limit is finite, then j is said to be stable.

We consider semirecurrent phenomena and provide some examples. The main result is that the semirecurrent set ζ corresponds to the range of a Markov renewal process (MRP) and conversely, a semirecurrent set can only arise in this manner. For details of the results from Markov renewal theory used in this paper see Cinlar ([4], Chapter 10). We construct a Markov subordinator with a unit drift whose range turns out to be a semiregenerative set. In the case where E is finite we prove that every semiregenerative set corresponds to the range of a Markov subordinator. Our approach yields results analogous to Kingman's ([7], Chapter 5) for quasi–Markov chains. While our approach (based on Definition 1) is thus more rewarding in these respects, our techniques are simpler, being based on properties of Markov renewal processes. Bondesson [1] has investigated the distribution of occupation times of quasi–Markov processes. We shall not investigate this problem for semiregenerative phenomena.

In the literature there are extensive investigations of semiregenerative processes. These are processes imbedded in which there is an MRP (or equivalently, a semirecurrent phenomenon). We take the view that semiregenerative phenomena are important by themselves and therefore worthy of study. In particular, the theory developed in this paper provides a proper perspective to the work of Kulkarni and Prabhu [8] and Prabhu [10].

REFERENCES

[1] Bondesson, L.: On occupation times for quasi–Markov processes. J. Appl. Prob. 18 (1981), 297–301.

[2] Cinlar, E.: Markov additive processes. I. Z. Wahrscheinlichkeits–theorie verw. Geb. 24 (1972), 53–89.

[3] Cinlar, E.: Markov additive processes. II. Z. Wahrscheinlichkeits–theorie verw. Geb. 24 (1972), 95–121.

[4] Cinlar, E.: Introduction to Stochastic Processes. Englewood Cliffs: Prentice–Hall, 1975.

[5] Feller, W.: An Introduction to Probability Theory and its Applications, Volume 1, Third Edition. New York: John Wiley, 1967.

[6] Kingman, J.F.C.: Linked systems of regenerative events. Proc. London Math. Soc. 15 (1965), 125–150.

[7] Kingman, J.F.C.: Regenerative Phenomena. New York: John Wiley, 1972.

[8] Kulkarni, V.G. and Prabhu, N.U.: A fluctuation theory for Markov chains. Stochastic Processes Appl. 16 (1984), 39–54.

[9] Neveu, J.: Une generalisation des processus a acroissements positifs independents. Abh. Math. Sem. Univ. Hamburg 25 (1961), 36–61.

[10] Prabhu, N.U.: Wiener–Hopf factorization of Markov semigroups–I. The countable state space case. Proceedings of the Seventh Conference on Probability Theory (M. Iosifescu, Ed.). VNU Science Press, Utrecht (1985), 315–324.

Invited Paper 15.1
46th Session of the ISI

Student:
A Tool for Constructing Consultation Systems in Data Analysis

William A. Gale
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ, 07974

1. Introduction

Knowledge-based consultation systems (often called expert systems) have become common in the last decade. Outside the field of statistics, several commercial systems have been built. Within statistics progress has been limited to methodological feasibility studies, beginning with REX (Gale and Pregibon, 1982; Pregibon and Gale, 1984; Gale, 1986b). Since then Muse (Dambroise and Massotte, 1986), Express (Carlsen and Heuch, 1986), and unnamed systems by Berzuini and others (1986) and Darius (1986) have been described.

I mention these first level consultation systems to distinguish Student from them. Student is more than a consultation system, since it is primarily a tool to help a statistician build such consultation systems. But since Student also serves as the vehicle for the constructed knowledge-based consultation systems, it includes the capabilities of the first level consultation systems.

Student is designed to allow a professional statistician to build a knowledge-based consultation system in a data analysis technique by selecting and working examples and by answering questions. The statistician does not need to know the internal representation of the strategy demonstrated, and does not need to know how to write a knowledge based program. He does need to be fluent in the underlying statistical system, a more natural expectation of a statistician.

REX is a working demonstration of the type of consultation that Student will provide. It allows a novice to use advanced regression techniques safely by systematically checking the assumptions of the techniques. It provides guidance to what tests need to be done and when, interpretation of the results of tests and plots, and instruction in statistical concepts. It has appeared that REX, while designed for use by novices, is interesting to expert statisticians, because it makes explicit much knowledge that has not been formalized. Most experts have also expressed interest in using such a consultation system because it automates many tasks that they know they want to do, but don't always do.

Like REX, Student is based on an underlying statistical analysis system, and constitutes an interface to that system. Student uses Quantitative Programming Environment, QPE. (Chambers 1986) as the underlying system. Briefly, QPE has been designed as a successor to S (Becker and Chambers, 1984). The external syntax and appearance have been largely maintained. But QPE was designed to be an *environment*, that is, to contain programming, browsing, debugging, and editing capabilities. The design of Student assumes that the statistician using Student to create a consultation system knows how to use QPE.

A methodological prototype study of Student (Gale 1986c) was built using Lisp and a Symbolics machine. The current version of Student is intended as a product definition study. It is programmed in the language provided by QPE, since this would be the most likely delivery language for a product. The goals of the QPE version are to study issues such as speed, usefulness to statisticians, and generality of the conceptual framework used by Student. This version is currently a partially developed system that has only begun to be used by statisticians. It has not yet begun to answer the product issues posed, but shows the knowledge acquisition methods more clearly than the prototype, and has begun to be used to acquire a few different data analysis strategies.

By using QPE, hardware and software requirements are minimized. QPE will run in most Unix™ environments. Wherever QPE runs, Student will run. Student is not a product, but if it were, it would require a machine with Unix, and QPE software.

This articale appeared in the Bulletin of the International Statistical Institute, Vol. 52, pp 1-18. Permission of the author and the editor of that journal to reproduce it here is appreciated.

What Student adds to the capabilities of REX is the capability to acquire its knowledge base by interview and demonstration. The demonstration approach was proposed by Gale and Pregibon (1984), and tested in the Lisp prototype (Gale 1986c).

The knowledge base used to conduct a particular method of data analysis has been called a *strategy*, and the term will be used here. Section 4 defines strategy. Briefly, a data analysis strategy includes knowledge about the kinds of problems that can occur in using the method, how to test for them, what to do if they occur, and how to communicate the problems and solutions to a novice user.

The importance of acquiring a strategy by interview and demonstration is considerable. In the current state of building knowledge-based consultation systems, two distinct roles, usually played by two different people, are standard. One is the role of subject matter expert, and the other is the expert in the inference engine used, or *knowledge engineer*. In building REX, I played the knowledge engineer, while Daryl Pregibon played the statistical expert. This procedure requires the knowledge engineer to learn a lot about the subject matter, or the subject matter expert to learn a lot about the inference engine and programming, or both.

Student's primary goal is to allow a statistician, who does not know how the inference engine is built, to build a knowledge based consultation system without the involvement of a knowledge engineer. This should support greater efficiency in building consultation systems in data analysis.

There is a substantial secondary benefit as well. A statistical consultation system will be used in many other ground domains, such as physics, psychology, or business analysis. Current AI techniques are not adequate to handle knowledge in multiple domains, so we built REX with the explicit assumption that the user was willing to learn statistics concepts and vocabulary. This assumption will be reasonable for many analysts, but it will be unreasonable for many managers or low frequency users of statistics.

Student provides the means to specialize the knowledge and vocabulary used to guide a consultation in data analysis. Because it can learn by interviewing a statistician using locally relevant examples, it can be provided with strategies shaped to local environments. This will increase the market size for a Student-like product as compared to a REX-like product.

Another significant benefit of removing dependence on a knowledge engineer is the capability to specialize a system to a local environment. When Student is first acquired by a group such as a quality engineering group, a specialist statistician can select examples from the group's files and work them in the Student environment. After this specialization training, the engineering experts would use Student for consultation, returning to the statistician with problems beyond its training. When such a problem seemed frequent, the statistician would work it as an addition to the strategy. If it seemed infrequent, then it would be worked by hand.

There have been three main challenges in building Student. First, the system had to support the acquisition of the *first* example. In a rule based system, the first rules to be acquired are typically different from later rules, because a rule based system uses a core of rules to encode control information. A subject matter expert would not be able to provide control information.

Second, Student had to acquire knowledge from a new example that was *consistent* with its previous examples. Consistency means that all the examples that the statistician considered as properly worked, remain so when the additions to the strategy are made.

Third, the system had to support deliberately inconsistent changes to strategies over a long period of time. Current technology, such as used for REX, results in a "compiled" strategy, which is difficult to change.

The current version of Student has made clear that the first two of these challenges have been met, and it suggests that the third can be met. These challenges have been met by the development of an artificial intelligence technique called *knowledge-based knowledge acquisition* (Gale, 1986e). Knowledge-based knowledge acquisition means restricting the domain of knowledge that can be acquired, and developing a conceptual model of the restricted domain.

Student is restricted to acquiring *data analysis* strategies. It is not a general purpose knowledge acquisition program for a general purpose inference engine. With this restriction, I have been able to provide a conceptual model for strategies of data analysis. For instance, we know we have to deal with data sets, and we have provided representations to deal with them. The conceptual model specifies that

the analysis consists of looking for violated assumptions, and if found, of finding a cure. It specifies that we look for violated assumptions by making tests and by showing the user plots. I derived this conceptual model by inference from REX, and by considering extension of the methods to other data analysis techniques. Having this conceptual model provides enough structure to guide the user through the first analysis of a given kind, and to acquire additional consistent examples. It is still a research question how far this view based on work with REX will generalize, and how well inconsistent changes can be treated.

Student is written in modules that fall into three groups: control, data structure management, and learning. The data structure management modules can be distinguished as managing primitive or composite structures. Student then acquires knowledge by filling in data structures, which become a significant part of the system. The control and data structure management modules are nearly independent of statistics knowledge. The learning modules are specific to data analysis. This paper will focus on the learning modules.

The ideas for Student were proposed by Gale and Pregibon (1984), and tested in a Lisp prototype by Gale (1986c). The current QPE version is a partially developed system intended to explore issues relevant to making a product. It is on the leading edge of the line of research applying artificial intelligence techniques in statistics for the benefit of statistically naive users (Gale, 1986a; Haux, 1987). Our statistics department has begun to use this version of Student to better define the requirements for a useful product.

2. The Appearance of Student to the Statistician

The appearance of Student to the statistician is partially conveyed by the following transcriptions of sessions with Student. The sessions show how a statistician begins to instruct Student. This phase demonstrates most clearly the knowledge built into Student, as opposed to the knowledge that it acquires.

2.1 Introducing a New Data Analysis Method

In the first session, work is begun on a strategy for a data analysis technique new to Student. This is the first stage of building a new consultation system. The example supposes that spectrum analysis (Bloomfield 1976) is to be defined for the system. The statistician's entries are shown by a ">" and boldface text. The computer's screen output appears in *italics*.

>student( )

Do you want to
1. get a consultation from me
2. instruct me in a strategy
>2

There is no currently selected analysis. Do you want
1. univariate description
2. simple regression
3. New Analysis
>3

Please type a word for the name of the new type of analysis (_____ analysis)?
>spectrum

At least one variable is mandatory for spectrum analysis.
What is the name of the first required variable?
>time.series

Each input variable will be copied to a standard QPE variable.
What standard name would you like to use for the time series variable?
>x

time series is now the required variable for spectrum analysis.
Please enter another name for another required variable,
or '.' (dot) if there are no more.
> .

There are no optional variables now. Please
enter a name for a variable, or '.' (dot) if there are none.
> .

Notes on the session to this point:

The session begins from within QPE by calling student just as any other QPE function is called. Since Student is built within QPE, it is easy to use Student for routine analyses, switching to the full power of QPE if the analysis becomes non-routine.

The first point that Student must establish is whether the user is a statistician who wants to instruct Student, or a consultation client. This is essentially a branch to two entirely different programs which share the same data structures. The data structures are acquired from statisticians and used for consulting.

Student then asks which analysis will be worked on. Any existing analysis may be modified by a statistician. Notice that the list of options provided is a dynamic one that depends on what kinds of analyses Student knows about. Also, the option to add a new kind of analysis is only shown if the mode asked for was student.

The *reason* we know that we must have at least one mandatory variable is that we are doing *data analysis*. There may be other mandatory variables, and there may be optional variables. This is a clear and simple example of how the restriction to data analysis provides information which enables knowledge acquisition. By organizing the information about data analysis as much as possible, we try to build a formal *cognitive model* of data analysis.

The session continued:

Analysis Revision Menu
1. select or define analysis (spectrum now)
2. create a new strategy (copy or construct)
3. define a new example
>2

There is no existing strategy to copy.
Please type your last name.
>gale

Please type a word to use for the strategy name
(or just a '.' (dot) to accept "gale" as a default).
> .

Analysis Revision Menu
1. select or define analysis (spectrum now)
2. create a new strategy (copy or construct)
3. define a new example
4. select a strategy
5. delete a strategy
6. stabilize a strategy
7. refine the gale strategy
>7

Strategy Refinement Menu
1. show strategy & statuses
2. select or define an example
>1

NA.in.inputs OK
final.calculation OK

Notes on the continuation.

Once an analysis is selected, the top level menu becomes the analysis revision menu. An analysis is needed to select the dynamic entries to the menu. The menu shows that with a minimally defined analysis, we can define a new strategy or a new example.

The session continued by defining a new strategy. The system records the name of the author of the strategy, and gets the date it is begun from the operating system.

Communication requires many names, and they have to be convenient for the people using Student. Thus, Student needs to ask a lot of names. Wherever possible, the system suggests a default, but the final choice is up to the statistician.

After creating an empty strategy, the analysis menu has expanded. Before there was a strategy, there were none to select or delete, so there was no sense offering these options. A strategy is "stabilized" to make it available for consultation. So long as a strategy is considered stable, it can be used for consultation, and it cannot be modified. This is just a reminder to the statistician, since it is simple to stabilize and destabilize, or to copy a stable strategy and modify the copy. But it is important that the statistician carefully consider which strategies Student will be allowed to use for consultation.

This session is ended as Student shows the two minimal features automatically created for any new strategy. QPE provides "NA" as the result when asked to take the logarithm of a negative number, or to divide by zero. Student infers from NA's in transformed variables that inappropriate mathematical manipulations have been made. The knowledge about the limits of the functions is thus distributed among the functions themselves. If it were not there, it could be provided as background knowledge, but there is no reason to duplicate the knowledge. This does mean that strategies need to specify what should happen if an original input contain NA's. The feature initially provided will detect NA in any input, but awaits learning what to do until some example triggers it. The other feature initially provided is an empty shell to place a final calculation and report into.

Before the strategy extension cycle can be begun, an example must be provided.

2.2 Introducing an Example

In this session, Student is shown the existence of a new example. No demonstrations can be made without examples, so this step is necessary to continue constructing a consultation system. The user has selected simple regression analysis (Mosteller and Tukey 1977) for this session, at a point just after defining simple regression and one minimal strategy named "basicsr."

Analysis Revision Menu
1. select or define analysis (simple regression now)
2. create a new strategy (copy or construct)
3. define a new example
4. select a strategy
5. delete a strategy
6. stabilize a strategy
7. refine the basicsr strategy
> 3

Please type a word to use for a short name for this example.
> **brain.body**

Please type a word to use for the response variable data set.

> brain.y

Please type a word to use for the explanatory variable data set.
> brain.x

Do you have a weight variable for this example?
> no

Please type a brief description of the brain.body example.
> The response variable is the average brain weight in grams,
the explanatory variable is the average body weight in kilograms,
for 62 terrestrial mammalian species. Data from Weisberg, p128.

Notes on this session:

The simple regression data analysis method was defined to have two required inputs and one optional input. The required inputs are called "response" and "explanatory", and the internal QPE names are "y" and "x". The optional input is called "weight" and its internal name is "w." This session shows how the information acquired by Student is put to use and becomes difficult to distinguish from the knowledge it starts with.

If the short name had been chosen as "brain," the system would have located "brain.y" as a data set named by concatenating the short name and the internal name of the response variable. It would have assumed that the data set was so named precisely to be used as the initial input for the response variable. It would likewise have found "brain.x" as a data set for the explanatory variable. As it is, the system has checked that the data sets of the given names exist. It then constructs code to assign these initial values to the data sets "y" and "x." It does not execute this code now, but stores it as part of the definition of the example.

The system did not find a data set named "brain.body.w", so it asks if there is a weight variable for this specific example. When it learns that there is no weight variable, it uses stored code describing how to generate default values for the weight variable. The code used was acquired by demonstration during the initial definition of the simple.regression analysis frame.

The description of the example is treated as unprocessed text. It is available to those modifying a strategy to see what examples the strategy was developed with. Asking for it is a reminder to the statistician that the information will be needed by others later. It is probably easier to give this information now than in the future. The reply given here shows that there is information that could be broken down and some of it made available to the machine. The meaning of each variable, their units, the sampling units, and the source of the data might each need to be asked individually.

2.3 Strategy Extension

This session shows the usual cycle for strategy extension. It begins with a minimal strategy for simple regression. I have shown this session without the full menus, only the menu line selected by the user as the user's input.

Analysis Revision Menu
> 7. refine the basicsr strategy

Strategy Refinement Menu
> 2. show examples and evaluations

brain - unanalyzed

Strategy Refinement Menu
> 3. select an example

The only example available is the brain example.
> 4. analyze the example

beginning to consider NA.in.inputs feature with no argument
beginning to consider final.report feature with no argument

Strategy Refinement Menu
> 5. REVISE strategy by inserting a new feature

Which feature is the last one correctly analyzed?
1. none of the below are correct
2. NA.in.inputs(none)
3. final.report(none)
> 2

Please type a word to use for a name for the new feature.
> skewness

Please tell me why skewness is important for simple regression.
> The skew points are unduly influential.

Please type a word to use for a name for this test (the ... test),
or just a '.' (dot) to accept skewness as a default.
> .

Please type a word to use for a QPE name for the test statistic,
or just a '.' (dot) to accept skewness as a default.
> .

Type 'return()' to make your last expression define skewness.
Student: **qtls<-qtl(y) #quartiles of y**
Student: **qtls**
(4., 17.25, 169.)
Student: **(qtls[3]-qtls[2])/(qtls[2]-qtls[1])**
11.4528
Student: **return()**

The value of skewness on this example is 11.4528.
What preliminary LOWER limit do you suggest?
> 1.5

What preliminary UPPER limit do you suggest?
> 3

The interpretation of the first test result is severe.
Is this your intention?
> yes

This test has just one input variable. This can be treated
as an argument if you want, but doing so will make the result
unavailable to further computation.
Do you want this to be a feature with an argument?
> yes

Please type a word to use for a short name for this transform (the ... transform).
> log

IP-15.1

I am setting up a temporary environment. Please show me how to make a log transform by providing code to redefine ALL NECESSARY input variables, ENDING with a redefinition of y.
Type 'return()' to make your last expression define y.
Student: **log(y)**
(3.79549, 2.74084, ...

Student: **return()**
You have shown me: expression(y <- log (y))
is this a satisfactory definition of the log transform?
> yes

The log transform will reduce the problem severity from severe to mild.
Is the log transform acceptable to you?
> yes

Committing to the log transform.

Strategy Refinement Menu
> 4. analyze the example

beginning to consider NA.in.inputs feature with no argument
beginning to consider skewness feature with argument y
making log transform
beginning to consider final.report feature with no argument

Notes on the session:

All strategies for a given analysis method share the same set of examples, each defined by specifying the input variables. Each strategy has its own records about how well the strategy has analyzed the example. Each example has status unanalyzed, acceptable, or unacceptable. To start with, an example is unanalyzed. After a strategy revision, all examples are marked unanalyzed.

After selecting and analyzing the brain example, it is found to be unacceptable, because there is no basis for declaring it acceptable. One action possible for an unacceptable example is to declare it is acceptable. Then it is so marked, and the pattern of transformations and their reasons (features of arguments) is stored. Any other analysis that makes the same sequence of transforms for the same reasons will be automatically marked acceptable. An acceptable example can be declared unacceptable, which causes the pattern to be stored as a known bad pattern.

The other options for an unacceptable example all revise the strategy. The session shows one way to revise the strategy, by inserting a new feature. Other ways include deleting a feature, and revising a feature. To insert a feature, we must know how far the analysis is considered correct. Then the new feature will be inserted so that it will be tested following the last correct feature.

The acquisition of a test shows the system collecting code to define the test. The statistician is in a slightly modified QPE environment, free to examine data known to Student, call on any predefined QPE functions, and to plot as may be useful. The modifications are that the user may not refer to data not known to Student, and may not make an assignment to a global variable known to Student. When the user types 'return()', a legal QPE expression with a special interpretation here, control returns to Student. The program then cleans up the series of expressions into a minimal set required to define the desired variable. In the example, the line on which the statistician examined the values of the quartiles will be deleted.

Student will infer lower and upper limits from the statistician's actions over many examples. But when there is only one example, the induction method fails. Therefore a set of preliminary limits is requested. Their importance declines as more examples become available. The preliminary limits can also be set by an automated Monte Carlo method, but it is too slow for interactive use.

The system examines the code produced, and finds that only one variable was used to define skewness. In such a case, generalization is frequently useful. It simplifies the process of reconstructing any given value to have such generalized functions not be used in further calculations. This appears to be acceptable in common cases where generalization is useful. If it is unduly restrictive, a more complex internal method can be programmed.

The system then asks for a demonstration of what to do if skewness is found to be a problem. A transform specifies each input's new value. The previous values of the inputs and intermediate results based on them are available for the new specification.

Student always creates a temporary environment when it considers a transform. The transform is made in the temporary environment, and the test for the feature is applied. If the result is still unacceptable, the transform is not committed, but the original environment is restored. This procedure is followed even on the time that Student is shown how to make the transform.

This completes the demonstration of the skewness feature. Student now works the example by making the log transform of the response variable. The next step will be to show it that the skewness of the explanatory variable needs to be examined. This will be much shorter to show, since the same feature can be reused with a different parameter.

3. The Knowledge Acquisition Method

3.1 A Critique of Knowledge Acquisition in REX

Developing a strategy for use in REX was a labor-intensive process. Two phases can be distinguished. In the first phase the statistician responsible for the strategy, Daryl Pregibon, chose a half dozen regression examples that clearly showed some frequent problems. He then analyzed them using interactive statistical software with an automatic trace. After analyzing the group of examples, he studied the traces and abstracted a description of what he was doing. We coded this as a strategy for REX and tried it on a few more examples. He revised the strategy completely at this point, and the second phase began.

In the second and longer phase, one of us would select one additional regression example and run REX interactively on the chosen example. Since we selected the example knowing what would stretch REX, REX usually reported a severe problem that it didn't know how to fix. Then we would modify the strategy so that the example would be handled. This process was iterated through about three dozen more examples.

Based on this experience, and on a feeling that it was typical of other techniques, we do not believe it is possible to build a data analysis strategy without working through many examples. One must make many decisions to build a strategy, and there is no literature simplifying the task. Therefore the only available defense of a strategy is to demonstrate performance, which requires working many examples more than those used to build the system. On the other hand, our experience also leads us to believe that it is easy to generalize from data analysis examples. The basis for generalization is usually a statistical test that statisticians can provide. Generalization then consists of determining the range of values of the test for which the demonstrated technique holds.

However, the way in which we worked examples for REX was far from ideal. The first difficulty with our method was assuring ourselves that a strategy modified to work one additional example still worked all previous examples. We could by brute force run REX in batch mode on all previous examples and see if the performance was the same. Usually we reasoned that most of the previous examples could not be affected, and checked the few that might be affected by hand. Naturally, the more examples worked, the more severe this problem became. The need to check consistency in batch mode for a system designed to be interactive reduced the flexibility of the strategy developed.

Second, the method used was the epitome of the currently standard two-person development of expert systems. I built the inference engine used while Daryl was responsible for the strategy developed. Whenever Daryl wanted to do something he hadn't done before, we had to huddle, as Daryl was learning a language he would only use to build one program. In a department with twenty professional statisticians and one person intimately familiar with the inference engine, it was not clear how many additional data analysis techniques could be handled by this two person approach.

Third, it would be difficult to modify the strategy in REX. Modifiability is important because a growing literature on strategy (Gale, 1986a; Haux, 1986) can be expected to suggest desirable changes. It is also important because users will probably want to modify strategies to their particular needs. However, the first two problems would make this difficult: to specialize the program a local statistician would have to learn a language used by no other program in the world, and the modifications made might inadvertently destroy some capabilities of the strategy.

However, the development of REX contributed greatly to following work. It provided us with the beginnings of a conceptual model for data analysis: a data analysis consists of a desired calculation, assumptions required for the calculation to be meaningful, tests for the violation of the assumptions, and transformations to ameliorate the violations. The classes of frames used in REX provided us with an initial list of classes of primitives that has remained useful and has been expanded into a fuller conceptual model of data analysis.

3.2 Knowledge Acquisition in Student

The necessity of working examples to build a data analysis strategy suggested the possibility of acquiring strategies directly through that process. A system should assist the teacher in establishing consistency across all examples worked, and should not force a statistician to learn an obscure language. It appeared that examples might provide a language suitable for communication between statisticians and computers.

The first issue encountered in designing Student was how to learn from the *first* example. In a system without knowledge, there is simply no basis for use of information provided in working an example. By providing Student with the conceptual framework induced from REX, we have built a system that can deal meaningfully with an example even when it has seen no previous examples. The rather limited use of code collection in Student shows how much of the knowledge it is acquiring is not knowledge that could be inferred from just watching the analysis of an example. Even for the parts heavily dependent on code, if the system did not have some notion corresponding to "plot", "test", and "transform", it would not be able to deal with code having these different functions. In short, understanding the first information provided is possible because the system is limited to data analysis, and because it has been possible to build a conceptual framework for data analysis.

The conceptual framework used in the current version of Student has the fifteen classes of primitives shown in the following table. Each instance of a primitive is represented by a frame. In the table, indentation shows that names of instances of the primitive indented occur as values in some slot of the superordinate primitive. That is, the relation shown by indentation is "uses information from."

- analysis
  - input variable
  - example
  - feature
    - test
    - plot
    - transform
    - report
  - strategy
    - linear
    - conditional
    - repeated
- concept
- class
- consultation

Each primitive has a set of slots, which are also chosen to reflect the structure of data analysis. As an example, a simple primitive is the input variable frame, which has only a few slots:

input variable
- *external name of input*
- *required or optional*

default if optional
data type
internal variable name

The content of the instances of these primitives is the information that a consultation system must have. For instance, when asking a consulting client for a specific input, it is necessary to know the common name of the input. Likewise, the system must know whether to insist upon having a given input variable before beginning the analysis (required or optional), and what default to use if the user does not have an optional input. The system must also know what data type the input requires to determine if submitted data is possible. Since we do not want to overwrite input data with later calculations, we need a standard variable name to copy the input to.

Knowledge-based knowledge acquisition in this context means specifying how the contents of each slot will be acquired. For the input variable primitive, each slot could be acquired by asking the teaching statistician. Most of them could also be acquired more actively. The internal name could be created from the external name and perhaps a unique number. Acceptable data types could be inferred from the data types of the inputs to the set of examples provided. Optional variables and their defaults could be inferred as those with repeated inputs. It seemed better in each of these cases to ask the teaching statistician and then use the information to check inputs to teaching examples.

Thus, specific techniques designed for the specific knowledge in each slot were chosen. Student uses four specific techniques: interviewing, limits induction, Monte Carlo learning, and background knowledge.

Most cases are handled by interviewing. Knowing what is needed, and having a statistician at hand, it is easy to just ask. Even so, exactly how to ask for the information varies between menus, fill in the blank, multiple simultaneous choice, and free response. And of course the prompts vary with the item.

Monte Carlo learning can establish initial notions of the distributions for test results. The distributions in turn can be used to set initial cut points, or *limits* for distinguishing severe, mild and insignificant cases of assumption violations.

Limits induction is inference of limits on test ranges from test results and action (transform) or non-action by the statistician. Let v_i be the value of a test on the ith data set, and a_i be T or F as the statistician acted or didn't act. Set the lower cut point as $\max(v_i \mid a_i=F)$ and the upper cut point as $\min(v_i \mid a_i=T)$. Then for test values above the upper cut point, the statistician has always acted, and for values below the lower cut point, the statistician has never acted. This simple scheme is slightly modified to include the Monte Carlo results.

Knowledge-based knowledge acquisition has several advantages. First, the information in each slot is necessary for a consultation program. Systematizing the knowledge to acquire from a statistician speeds construction because the system won't forget what is needed.

Another advantage of knowledge-based knowledge acquisition can be shown in the acquisition of an input variable. It is almost always appropriate to run a number of tests on each input variable by itself. Without knowledge-based knowledge acquisition each time a new variable is given, a battery of tests must be specified by the teaching statistician. However, it is easy to keep track of what tests have been used for all input variables by data type, and to suggest these to the statistician. Since the tests are based only on knowing the data type of the input, they will often be appropriate in many different data analysis procedures. The domain knowledge we are using here is that the tests are similar in many different analysis types, and that they are reasonably organized by data type.

As another example, a statistician may notice after some time of programming that an optional input variable is possible. One would then back up and increase the generality of numerical procedures to accommodate the extra variable. With knowledge-based knowledge acquisition, the statistician is encouraged to think of optional inputs at the beginning of the construction process, thus avoiding the costs of reprogramming. This encouragement may not always be effective, but it can only work in the direction of reducing the problem. In short, by providing a framework for data analysis, the statistician is encouraged to think in previously successful terms.

Acquiring first examples does not address all the problems in building a knowledge acquisition system. However, the domain restriction has been useful for extending a given body of knowledge as well as

beginning it. Extension of knowledge for a given data analytic technique involves demonstrating more assumptions, how to detect their violation, and how to fix them. The same techniques used for initial acquisition suffice here. However, it is also necessary to check consistency for previously worked examples.

Knowledge-based knowledge acquisition has also been useful for dealing with consistency as the number of examples and the strategy have grown. Consistency means that after incorporating information on a new assumption, the recommended analyses of all previously worked examples are not changed. This is a requirement analogous to logical monotonicity. Some changes can be proved consistent by using domain knowledge. The domain knowledge consists of a theorem, and the proof consists of verifying the hypotheses of the theorem, so this is not automatic theorem proving. The proof may use data that could be specified and collected when the previous examples were demonstrated. This will be more efficient than rerunning examples. Other cases, such as showing that a new test is not passed for an old example, require new calculations. Domain knowledge is able to specify data to save that will make such checking faster than completely reworking an example.

Of course, the check may find that a change is inconsistent. That is, that the recommended analysis for at least one previous example has changed. Then the statistician will need to *revise* the existing body of knowledge. This might just consist of blessing the revised analysis for the inconsistent examples. Or it may require revising the strategy, perhaps revising the assumption just added. This can be assisted by domain knowledge encoded as editing procedures.

3.3 A Critique of Knowledge Acquisition in Student

Interviewing is useful. A knowledge-based interview is easy to write, since one knows exactly what to acquire. Interview procedures attached to slots are easy to keep track of, so that it is easy to see if all slots can be acquired.

A research issue is how much can and should be acquired by interviewing, and how much must or should be provided as initial knowledge. The Lisp prototype tested this by attempting to acquire everything by interviewing. It appeared that everything *could* be acquired this way. However, experience with this extreme approach led to deciding to provide some items as initial knowledge. The collected reasons used to justify initial provision of an item were

(1) distractingly frequent requests for information,
(2) richly structured information,
(3) stable and non-controversial information.

For example, data types (vectors, matrices, time series, ...) are being built in for reasons 2 and 3. An initial core of technical definitions will be provided for reasons 1 and 3.

The original idea of programming through demonstration of techniques on examples needs further development. In the Lisp version of Student, demonstration of examples seemed slow and clumsy. As Student has developed, the settings in which demonstrations occur have been restricted to key points about a particular example, so that the demonstrations become short sequences in a well understood setting. This has helped, and it is useful when describing a plot or test to have an example to do the operations on immediately. However, the process is still not flexible enough to allow exploration and final selection of one of several approaches tried. The statistician needs to approach the system with a clear idea of what will be demonstrated. There is, however, key information in the examples and I believe the current system is a useful start towards a more flexible system.

We found in building REX that the most powerful explanations in statistics were not verbal, but graphical. Thus we programmed before and after plots for each transformation. Student is able to make these automatically from plots acquired while being shown how to detect an assumption violation. This is a convenience.

Monte Carlo learning seems like a technique with much wider applicability for statistical systems to learn about statistical tests. Its use will be limited to overnight applications.

Limits induction is apparently a useful idea. It can describe what a statistician has actually done, possibly pointing out a poorly worked example, or a poor test. It can be used to alert statisticians to taking an action that is not consistent with previous actions, but can be changed easily if they insist.

4. Statistical Strategy Representation in Student

4.1 Goals for Representation of Statistical Strategy

This section discusses what is meant by statistical strategy, how strategy is being used, and why it is being studied. The purpose is to derive the goals that must be met by representations of statistical strategy.

The term *statistical strategy* has been used to denote integrating previously known tests and transformations into coherent total approaches to data analysis. Although the term was suggested by in 1981 by Chambers, there is as yet no generally accepted definition of this term. Daryl Pregibon and I (1982) suggested that strategy would answer questions such as

"What do I look for?"
"When do I look for it?"
"How do I look for it?"
"Why do I look for it?"
"What do I have to do to look for it?"

Wayne Oldford and Steve Peters (1986) wrote "The term 'statistical strategy' will be used here to label the reasoning used by the experienced statistician in the course of the analysis of some aspect of a substantive statistical problem." David Hand (1986) stated "*statistical strategy* has been defined as a formal description of the choices, actions, and decisions to be made while using statistical methods in the course of a study." These definitions give the general flavor of the subject matter beginning to be addressed and for which representations must be sought.

A more informative view of what strategy must mean can be derived by examining the situations in which we want to use it. To this end, I would like to review two views of the data analysis process that have been proposed by Hand (1986) and Oldford and Peters (1986). Hand discussed four *stages* of analysis, while Oldford and Peters distinguished four *levels* of strategy. That is, Hand was concerned with entities which take place at different times, while Oldford and Peters' description is a classification.

Hand's four stages are (1) formulate aims, (2) translate into formal terms, (3) numerical processing, (4) interpretation. These stages were given specifically as stages in a multiple analysis of variance (MANOVA), but they appear to me to be general. The first stage is concerned with what dependent and independent variables are involved, how they are related, and what questions the researcher wants to explore. It is largely phrased in the language of the ground discipline. The second stage results in the translation from a problem statement in the ground discipline to a problem statement in statistics terms. The third stage consists of estimation, testing, data cleaning, and transformation. This stage functions within the statistician's language. The fourth stage consists of translating back to the ground domain. As Hand points out, there will be various loops in an analysis, returning to earlier stages to alter decisions.

Oldford and Peters suggest 'operational level' as a scale for thinking about procedures. At the lowest level are standard numerical procedures of statistics, such as least squares fitting or robust fitting. Selections from this level constitute the minimal components of a statistical package. Just above this level are such sub-procedures as collinearity analysis and influential data diagnosis. Each of these presupposes the existence of procedures in the layer below it. Above this layer lies a layer of techniques, such as regression analysis, spectrum analysis, or analysis of variance. The top-most identifiable level has strategies for analysis and for design.

The levels idea rests on a notion of a procedure using other procedures as building blocks to carry out its goals. The notion of stages is that of what is done first. The relationship between them is that the high level strategies are used first and more frequently. The low level strategies are used later if at all. Thus the higher levels of a hierarchy of techniques will correspond to the preliminary stages of a study.

4.2 Intentions in Studying Statistical Strategy

One intention in studying statistical strategy is clearly to respond to the programming opportunities available. All the programs discussed in the introduction can be said to have as their goal to help people choose statistical methods. This will require research by statisticians about how one should choose statistical methods. The strategy representation then should be usable by statisticians in communicating among themselves.

Since the current uses of strategy are for programs, the representation must be interpretable by machine.

The assumed users of all the programs cited appear to be untutored in statistics. Therefore, it will be important to interpret the numerical statistics in English. The strategy representation needs to ease preparation of reports on what has been done.

Implicit in the choice of technique and application of technique uses is the opportunity to assist users in many different techniques. The representation must then be capable of expressing how to make the required choices in many different data analytic techniques.

Another possible use of strategy is for statistical education. By clarifying what features the various tests and plots are designed to detect, when various features should be sought, and how to respond if they are found, it should be possible to educate students more effectively. A representation suitable for education may be considerably different from one for consultation, based on Clancey's experience with Guidon (Clancey 1984). Without a setting in which to test this use, the requirements are unclear.

The goals that emerge for a representation for statistical strategy are that it should serve as a communication medium between expert statisticians, students, and machines. It should be sufficiently expressive for strategies in the range of data analytic techniques. The machine uses include both deciding what to do and reporting why.

4.3 The Feature/Imperative Representation of Strategy in Student

This section describes the strategy representation evolved through REX and Student. Another representation is described by Gale and Lubinsky (1986), which compares the two representations.

The statistical knowledge in Student is represented by a symbolic network. The lowest level of this network consists of such things as strings representing commands to the statistical language, strings of English text to show the user, numbers representing limits for interpreting tests, and lists of past results. These lowest level entities are grouped into entities that represent such things as tests, plots, report fragments, and transformations. These are in turn grouped to represent what we call *features*, and the features are combined into strategies. This representation can be readily seen to correspond closely to Oldford and Peters' description of strategy by levels, although the contents of the lower levels are different.

Features represent statistical concepts such as outliers, mean, granularity, heteroscedasticity, and symmetry. When a statistician examines a strategy used by Student, features are the lowest level exhibited in the graphical presentation. When the Student program examines a strategy, it interprets the same structure as a set of commands, or imperatively. Thus I have called this representation scheme "feature/imperative." When interpreted imperatively, the strategy directs the program through a series of stages, analogous to Hand's description, but much more restricted in scope.

The feature/imperative representation has evolved through development of REX, and the prototype study for Student (Gale 1986c) to the current design. REX made two major contributions to following work. The first was a viewpoint for thinking about data analysis as a diagnostic problem. Briefly, one should list model assumptions (analogous to possible diseases), test the data set at hand for violations of the assumptions (analogous to symptoms), and if found select a transform of the data (analogous to treatment). The success of this approach depends on the representation of statistical knowledge. This was the second major contribution of REX. REX had a set of statistical primitives including tests, plots, assumptions, and transforms, which could be built with artificial intelligence techniques such as frames with slots, or objects with attached methods.

Features, plots, tests, and strategies are entities with enough usefulness as concepts that it is also useful to establish analogous entities in writing a program. The programing device used to represent these entities is called a *frame*. A frame is in the first place a place to store information. Named *slots* specify which information can be stored in the frame. Different types of frames are distinguished by what information will be stored in them. The bare bones of the strategy representation can then be stated by describing the types of frames, or primitives, used and what information is kept for each of them.

The Student prototype built on the insight gained from REX, and increased the number of primitives to ten. The current design for Student uses most of the primitives from the prototype plus a few more, as listed in the section 3. Descriptions of the primitives follow.

The concept primitive keeps information about technical statistical words. The purpose is simply to be prepared to define them for users. The more this definition can be tutorial, the better. This is the only primitive not used directly in the strategy.

The data type primitive keeps information about vectors, matrices, upper right triangular matrices, etc. There is a small collection of data types with a hierarchical structure. It provides information such as how to verify that a data set is of the required type, and how to generate a random example for a Monte Carlo study (Gale and Lubinsky, 1986).

The analysis primitive reflects that Student will handle several analysis techniques, such as regression analysis, description of univariate data, spectrum analysis, and analysis of variance. The analysis frame will show how many input variables are required, and how many are optional. It will also show what strategies are available. The input variable primitives specify such things as name, data type, and default value.

A strategy is validated by the examples that it works, and it is partially derived automatically from examples. Therefore each strategy will deal with a group of examples, each represented by an example primitive. The remainder of the primitives are used to express the strategy as a structure built of features.

The feature, test, plot, and transform primitives originated in REX and have been used in each system since. They describe how to test for a feature, hoe to show it to a user, and if its presence violates an assumption, what transforms can be considered to alleviate the problem. The report fragment primitive has been added to help generate a report. It seems likely to be elaborated.

The preceding discussion described how strategy in a broad sense is represented in Student. A strategy in the narrower sense of the strategy primitive is described formally as a combination of features. The combination used in Student is a programming language restricted by requiring a simple graphical display of an expression in the language. This is based on a decision to encourage statisticians to think about strategy by providing a vivid representation of a strategy. The restriction does not limit the strategies that can be described, but it may make a description clumsy. In interactive use only the graphical language is seen by the statistician. However, the formal language underlying the graphical expression gives it a clear definition of its meaning. It may also be useful as an off line recording and communication medium.

The language used is formally described as follows:

```
strategy = item (strategy / empty)
        ;
item = feature
        / 'if(' feature ')' (strategy ( 'else' strategy / empty )
                        / 'else' strategy )
        / 'for(' feature ')' strategy
        ;
feature = test-feature
        / strategy-feature
        ;
```

Informally, this is read that a strategy consists of a list of items. Each item is either a feature, a conditional strategy, or an iterated strategy. A feature is either a test feature or a strategy feature. A conditional strategy is a test on a feature, with one or two alternative strategies to consider depending on the test. A conditional strategy is a repeatedly tested feature with a strategy to consider whenever the test is passed.

The symbols of this language are given meaning by considering each feature, item, and strategy to be a predicate having value present or absent. A test-feature (a feature primitive) contains a test that can be applied to any example and a means of interpreting the test result to state that the feature is present or absent. This is the "ground truth" on which the language builds. A strategy is present if and only if at least one item is present. A strategy-feature has a strategy, and is present if and only if the strategy is present. A feature is tested according to its type, test-feature or strategy-feature. A conditional strategy is present if and only if the selected strategy is present. An iterated strategy is present if and only if the feature is present at least once and the strategy is present at least once. The feature of an iterated strategy must have exactly one argument that takes integer values starting with one. The iteration is

performed over successive values of the argument and terminates when the feature is not present.

This language can be diagramed using a node for each item. The details are given by Gale and Lubinsky (1986). Examples of the use of this notation for a strategy for unordered univariate description and the strategy used by REX are given there.

My belief is that this forms an easily learned language for statisticians, that it forms a sufficiently expressive language for data analysis strategies, and that it can be easily used by a machine to analyze data and report on the findings. All these points require further experience before the language is suitable for a product.

5. Prospective

Key questions still need to be answered before a reliable and easy to use program for building consultation systems will be available as a product. It is still not clear how far the conceptual model provided in Student will generalize, or how far it can be made to generalize. It is not clear how easy Student will be to work with, or how suitable the interface for statisticians is. The most fruitful avenue of continued research would appear to be to focus on statistical strategies, using Student to develop and compare strategies in commonly used data analysis techniques. We need experience with statisticians building strategies using Student and with consultations done using those strategies. This experience will show us what the opportunities are for further artificial intelligence applications.

REFERENCES

Becker, R. A., and J. M. Chambers (1984). *S: an Interactive Environment for Data Analysis and Graphics*. Wadsworth, Belmont, California.

Berzuini, C., G. Ross, and C. Larizza, "Developing Intelligent Software for Non-Linear Model Fitting as an Expert System." In *COMPSTAT 1986: Proceedings in Computational Statistics*, F. De Antoni, N. Lauro, and A. Rizzi, eds., Physica-Verlag, Vienna. pp 259-264.

Bloomfield, P., (1976), *Fourier Analysis of Time Series: An Introduction*, Wiley, New York.

Carlsen, F. and I. Heuch, (1986). "Express - An Expert System Utilizing Standard Statistical Packages." In *COMPSTAT 1986: Proceedings in Computational Statistics*, F. De Antoni, N.Lauro, and A. Rizzi, eds., Physica-Verlag, Vienna. pp 265-270.

Chambers, J. M., (1981). "Some Thoughts on Expert Software." In *Proceedings of the 13th Symposium on the Interface*, pp 36-40.

Chambers, J. M. (1986). "A Computing Environment for Statisticians." Presented at American Statistical Association annual meeting, Chigago, IL.

Clancey, W. J. (1984). "Use of MYCIN's Rules for Tutoring." In *Rule-Based Expert Systems*, B. G. Buchanan and E. H. Shortliffe, eds., Addison-Wesley, Reading, Mass.

Darius, P. (1986). "Building Expert Systems with the Help of Existing Statistical Software." In *COMPSTAT 1986: Proceedings in Computational Statistics*, F. De Antoni, N.Lauro, and A. Rizzi, eds., Physica-Verlag, Vienna. pp 277-282.

Dambroise, E. and P. Masssotte, (1986). "Muse: An Expert System in Statistics." In *COMPSTAT 1986: Proceedings in Computational Statistics*, F. De Antoni, N.Lauro, and A. Rizzi, eds., Physica-Verlag, Vienna. pp 265-270.

Gale, W. A., ed., (1986a). *Artificial Intelligence and Statistics*, Addison Wesley, Reading, Massachusetts.

Gale, W. A. (1986b). "REX Review." In Gale (1986a) pp 173-228.

Gale, W. A. (1986c). "Student Phase 1 - A Report on Work in Progress." In Gale (1986a) pp 239-266.

Gale, W. A. (1986e). "Knowledge-Based Knowledge Acquisition for a Statistical Consulting System." In *Proceedings of the Knowledge Acquisition for Knowledge-Based Systems Workshop*, J. H. Boose and B. R. Gaines, eds., Banff, Canada, pp 15-0 through 15-9.

Gale, W. A. and D. Lubinsky, (1986). "A Comparison of Representations for Statistical Strategies." In *Proceedings of Statistical Computation Section*, ASA, Washington, D. C.

Gale, W. A. and Pregibon, D. (1982). "An Expert System for Regression Analysis." In *Proceedings of the 14th Symposium on the Interface*, Ed. Heiner, Sacher, and Wilkinson, Springer-Verlag, New York. pp. 110-117.

Gale, W. A. and Pregibon, D. (1984). "Constructing an Expert System for Data Analysis by Working Examples." In *COMPSTAT 1984: Proceedings in Computational Statistics*, T. Havranek, Z. Sidak, and M. Novak, eds., Physica-Verlag, Vienna. pp 227-236.

Hand, D. J., (1986). "Patterns in Statistical Strategy." In (Gale 1986a) pp 355-388.

Haux, R., ed., (1986). *Expert Systems in Statistics*, Gustav Fischer Verlag, Stuttgart.

Mosteller, F., and J. W. Tukey, (1977), *Data Analysis and Regression*, Addison-Wesley, Reading, Mass.

Oldford, W. and Peters, S., (1986). "Implementation and Study of Statistical Strategy." In (Gale 1986a) pp 335-354.

Pregibon, D. and W. A. Gale (1984). "REX: an Expert System for Regression Analysis." In *COMPSTAT 1984: Proceedings in Computational Statistics*, T. Havranek, Z. Sidak, and M. Novak, eds., Physica-Verlag, Vienna. pp 242-248.

ABSTRACT

Student is an expert statistician's tool for building consultation systems in data analysis. To use Student, the statistician selects a technique of data analysis and choses examples for which the technique is appropriate. The statistician then demonstrates to Student how the chosen data sets should be analyzed. Various learning techniques are used by the Student program to build a *strategy* for the data analysis technique. These include asking questions, inference, Monte Carlo learning, and background knowledge. Student tests consistency between demonstrated examples and the evolving strategy. The statistician can change either the acceptable method for working an example or the strategy if the two are inconsistent.

Student is built within the Quantitative Programming Environment, a new generation statistical system. Use of Student only requires that the statistician know how to use QPE; no other language is needed. Student is being used to build strategies for univariate description, simple linear regression, and spectrum analysis.

The key artificial intelligence technique used to build Student has been called *knowledge-based knowledge acquisition*. This means restricting the domain for which knowledge can be acquired (to data analysis), *and* providing a conceptual framework for the domain. The conceptual framework for data analysis is expressed as a set of primitives representing such statistical concepts as strategies, features, plots, and examples. A strategy is represented as a network of frames each of which is an instance of one primitive.

RESUME

Student est un outil expert utilisé par les statisticiens pour construire des systèmes de consultation pour l'anayse de données. Pour utiliser Student, le statisticien choisit une technique d'analyse des données et des exemples pour lesquels cette technique est apppropriée. Le statisticien démontre ensuite au Student comment les bases de données choisis devraient être analysées. Des technique d'apprentissage diverses sont utilisées par le programme Student pour construire une *statégie* pour la technique d'analyse des données. Ces méthodes comprennent poser des question, la déduction, l'apprentissage Monte-Carlo et les connaissances de base. Le Student teste le cohérence entre les exemples démontrés et la statégie en cours. Le statisticien peut changer soit la méthode appropriée pour résoudre un exemple, soit la stratégie si les deux sont en contradiction.

Le Student fait partie de l'Environnement de Programmation Quantitative (Quantitative Programming Environment), un système statistique de nouvelle génération. Pour utiliser Student, le statisticien n'a besoin que de savoir utiliser le QPE; aucun autre langage n'est nécessaire. Student est utilisé pour développer des stratégies de description univariée, de régression linéaire simple et d'analyse de spectre.

La technique-clé d'intelligence artificielle utilisée pour réaliser Student a été nommée *acqisition de connnaissances baseé sur les connaissances*. Ceci veut dire limiter le domaine sur lequel des connaissaince peuvent être acquises (pour l'analyse de donnée), et fournir un cadre conceptuel pour ce domaine. Le cadre conceptuel pour l'analyse de données s'exprime sous la forme d'une base d'opérations des concepts statistiques tels que des stratégies, des fonctions, des tableaux, et des exemples. Une stratégie est représenté par un réseau de cadres dont chacun est un exemple d'une opération.

ON THE USE OF FACTOR ANALYSIS AS A PREDICTION TOOL

Oskar M. Essenwanger
U. S. Army Missile Command
Research Directorate
Research, Development, and Engineering Center
Redstone Arsenal, AL 35898-5248

ABSTRACT: Factor analysis is generally considered as being a diagnostic tool in statistical analysis. Since the mathematical background for factor analysis and the computation of empirical polynomials is the same, factor analysis can be useful as a prediction tool.

Factor analysis is compared with ordinary regression analysis as a prediction tool and some advantages utilizing factor analysis are discussed. In regression systems the individual terms are not necessarily independent while the factors are orthogonal. Predictors which have a time occurence later than the time of prediction cannot be included into regression systems but can be utilized in factor schemes. Furthermore, extreme values are usually underestimated in regression systems. Thus factor analysis may fare better especially for predictands whose frequency distributions are U-shaped rather than bell-shaped.

It will be demonstrated that prediction of ceiling height and cloud amount are two atmospheric parameters which may be predicted better with factor analysis than with a regression system.

1. INTRODUCTION. Many statisticians consider factor analysis as a diagnostic tool and prefer ordinary regression analysis techniques for predictions. One of the reasons may be the simplicity of the regression scheme. In addition, the availability of "canned programs" found today even for the small microcomputers (P.C.) contributes to this easy handling. However, regression analysis has some deficiencies which apply to factor analysis to a lesser degree. E.g. a new set of coefficients must be calculated for every added or omitted predictor. It is also known that predictors are not always independent from each other but the factors in factor analysis are othogonal. Thus a smaller number of factors (predictors) can achieve the same amount of residual (error) variance as in regression analysis.

Factor analysis is related to empirical polynomials whilch have been used in predictions. Consequently factor analysis is a prediction tool. In addition, two other facts are presented here which may favor the use of factor analysis as a prediction tool. Is is well known that regression analysis is based largely on persistence. If values of a parameter within the prediction interval are switching from a large positive deviation from the mean to an extreme negative departure or vice verse the regression model will fail to account for this variation. Furthermore, only those predictors known at the time of prediction can be included into regression analysis. In turn, factors can be derived from any set of predictors including elements whose value will not be known at the prediction time.

It will be illustrated in the subsequent sections that for prediction of ceiling height, cloud cover, or visibility, the factor analysis as a prediction tool may be b e tter suited than regression techniques.

2. MATHEMATICAL BACKGROUND. The regression model is based on:

$$(Y - \bar{Y})/S = A_1 (X_1 - \bar{X}_1) + A_2 (X_2 - \bar{X}_2) + \ldots + A_n (X_n - \bar{X}_n) \qquad (1)$$

In the factor analysis we can write:

$$(Y - \bar{Y})/S = B_1F_1 + B_2F_2 + \ldots + B_mF_m \qquad (2)$$

where m<n. In the notations above Y is the predictand, X_i are the predictors, F_i the factors, A_i, B_i are coefficients, and s or S denotes the standard deviation.

Examples for eqn (1) are given below for a prediction model of ceiling height:

$$Y_D = Y-\bar{Y} = 19Z_1 + 0.4Z_2 - 0.4Z_3 - 1.7Z_4 + 2.6Z_5 + 0.1Z_6 - 79Z_7 \qquad (3)$$

The predictors ($Z_i = X_i - \bar{X}_i$) in this model are Z_1 = visibility, Z_2 = zonal windspeed, Z_3 = temperature, Z_4 = relative humidity, Z_5 = surface pressure, Z_6 = ceiling height, and Z_7 = sky cover with clouds. Three forecasts for particular days follow where the subscript of the Y indicates the hour of the day. In the first case Y_{11} = 999 (synoptic code) at 11^h on a particular day at Stuttgart (Germany), and Y_8 = 999 at 08^h on this day. The predicted value from eqn (1) was 984 which is very close. On the second day Y_{11} was again 999 but Y_8 = 20. The predicted value for Y_{11} in this case was 125 which reflects the trend correctly but misses the magnitude of the change. Another example of a missed prediction is a case where the ceiling height dropped rapidly within 3 hours. Y_8 = 999, Y_{11} = 100, predicted 736. Again, the trend is consist but the magnitude of the change is missed. It will be illustrated later that the factor model in these cases of rapid change would have rendered a better prediction.

3. CLIMATOLOGICAL BACKGROUND OF PREDICTANDS. Before the factor model is presented we may inspect the frequency distributions of ceiling height, cloud amount and visibility (Fig 1-3). It is obvious that all three predictands do not conform with a bell-shaped distribution where extremes have a low probability of occurrence (e.g. $\pm$ 3 sigma = 0.27%). The other important fact is found in a survey of changes of the value of the element within a short time interval, here 08 AM to 11 AM (Table 1). In the last column of Tables 1A, B, C the change from one side of the mean value (indicated by the double bar) to the other side is summarized. We notice a change in 14, 9 or 18% for ceiling height, cloud amount, and visibility, respectively. In these cases incorrect predictions by the regression technique comprise a considerable amount of the total data. In addition, these cases of rapid changes may be of particular interest to the forecaster.

4. FACTOR MODEL. In this pilot study the first step of the factor model is a factor analysis whose structure matrix is displayed in Table 2. (For technical details see Essenwanger, 1986, 1987a, b, c) We deduce from Table 2 that factor one is highly related to ceiling height and cloud amount at 08 AM (GMT) but also to ceiling height and cloud amount 3 hours later. Unrotated factors and rotated factors differ very little for the first two factors which are the most important ones (see Essenwanger, 1987a).

The next step is the study of the factors. Table 3 exhibits the mean factors by ceiling height groups as an example. While factor one has a numerical value of - 8.22 when the ceiling height remains at 999 for the 3 hour time interval the value changes to - 2.40 when the ceiling rises from <50 to 999 (code in 100 ft). The following predictions cover the two cases where prediction by the regression model failed. In the first case a lifting of the ceiling height from 20 to 522 is calculated while the actual value is 999. This is a significant improvement over the number of only 125 from the regression model. In the second case where the ceiling drops from 999 to 100 the factor model renders 490 versus 736 from the regression model. Again, a significant improvement is obtained.

The predictions from the factor model, although considerably better than from the regression model, may not satisfy some skeptics. It must be stressed that these forecasts are based on mean factors, and better models may be developed given time and effort. This is only a pilot study. The real factors on these individual days would have resulted in the prediction of the precise observed value but even the utilization of mean factors was better than the forecast from the regression model.

5. MODEL COMPARISON. While these individual cases prove that a better prediction with the factor than the regression model could have been made in those particular cases it is necessary to study a larger data sample. Table 4 provides a decision tree from observations of ceiling height and cloud amount at 08 AM to derive the predicted value of ceiling height and cloud amount at 11 AM. The numbers of Y_1 and Y_4 were based on the mean factors such as in Table 3 leading to prediction as shown in Table 4. These factors had been derived from a data sample of N = 200 for Stuttgart (1946-1952) in January with a structure matrix as displayed in Table 2. The squared deviation between predicted values from Table 4 and actual values were summed up and divided by N and the variance. The results are disclosed in Tables 5A, B, C, converted to percentage.

The first column provides the results for the assumption that the value of the element is the same at 11 AM as at 08 AM (persistence). The second and third column lists the residual variance for one and four factors, respectively. Finally, the percentages in column 4 are given for the regression model, utilizing the observed value of the 7 elements at 08 AM without inclusion of the ceiling height, cloud amount or visibility at 11 AM. The latter 3 values would not be available at prediction time 08 AM but can be included into the derivation for the factor model.

Inspection of Table 5 reveals that the residual variance for the factor model is significantly lower than for the model based on persistence or the regression model. In fact, the application of the F-test proves a statistical significance above the 97.5 level (for N = 50 the threshold is 1.72, while for N = 200 the 99% value is 1.39 for the variances ratio, e.g. Hald, 1952). Table 5A displays the residual variances (in %) for the three predictands from models derived for this data set N = 200. Since we learn from Figure 1 that a data gap between 300 and 999 exists. One may suspect an excessive influence of missed extreme values. Therefore, consideration was given to convert all 999 values to 400 in order to reduce the magnitude of the variance and deviation from the mean for extreme values. As can be seen from the row "CEIL 2" in Table 5a the percentage figures have changed very little. Thus the data gap has little to do with the demonstrated improvement over the regression model by the use of a factor model.

It may be argued that the results should be favorable because the coefficients and factors have been derived for this data sample of N = 200. Thus an independent sample of N = 50 has been studied. The results are depicted in Tables 5B and C. Two versions were investigated. First (Table 5B) the coefficients for the models from the data set of N = 50 were derived and the same calculations as exhibited in Table 5A were performed. This computation reflects the "ideal case". It permits us to evaluate the degradation which is introduced by utilizing coefficients and factors derived from a different data sample such as the data of N = 200. Table 5C shows that the regression model experienced a larger increase of the residual variance than the factor model evidenced by the increase of the ratio REGR/F_1 from Tables 5B to 5C.

The critical observer may notice that the percentage for the residual variances are also changed for the persistance model from Table 5B to Table 5C. It may appear as a discrepancy at first but it can be explained. The variances in the 200 data sample are not identical with the variances in the 50 data sample. Consequently the percentage values change for Table 5C in accordance with the differences of the variances. It may be assummed that given a large enough

sample for Tables 5A and 5B this effect would disappear. This effect does not alter the basic conclusion that the factor model has provided better predictions than persistence or the regression model.

It may be of interest that the factor model based on 4 factors (Table 5C) did not render much improvement over a single factor model although for ceiling height and cloud amount the usage of 4 (mean) factors indicates a decrease of the residual variance (Tables 5A and B). Whether this is a sign of a general trend or a peculiarity of this special data set remains to be seen. Nevertheless, the one factor model in this pilot study led to a smaller residual variance than the 7 parameter regression model.

6. CONCLUSIONS. In predictions of atmospheric parameters such as ceiling height, cloud amount, and visibility, a model based on factor analysis may be better suited than a regression model. This may be due largely to the possibility to include predictands into the derivation of the factor model. A factor model has also an advantage that only one set of coefficients must be derived for the task of developing models for several simultaneous predictands. The results of this pilot study indicate a real potential of factor models in certain atmospheric predictions.

7. REFERENCES:

Essenwanger, O. M., 1986 - Comparison of Principal Components and Factor Analysis Method for Climatological Parameters. Proceedings of the Third International Conference on Conference on Statistical Climatology, Vienna, p. 40-45.

Essenwanger, O. M., 1987a - On Rotation in Factor Analysis of Atmospheric Parameters, Proceedings of the Thirty-Second Conference on the Design of Experiments in Army Research, Development and Testing, ARO Report 87-2, p 59-76.

Essenwanger, O. M., 1987b - Factor Analysis and Prediction of Cloud Parameters, Proceedings, 5th Cloud Modeling Workshop. 227-239

Essenwanger, O. M., 1987c - Prediction of Cloud Parameters by Using Factor Analysis Preprints, 10th Conf. on Probability and Statistics, AMS p. 43-46.

Hald, A., 1952. Statistical Tables and Formulas. Wiley & Sons, Inc., New York, pp 97.

TABLE 1: CONTINGENCY TABLE OF CHANGES OF ELEMENT IN PREDICTION INTERVAL

(STUTTGART (F.R.G.), JANUARY 1946-1953, N = 250)

A) CEILING HEIGHT (IN FEET)
11 AM

8 AM GMT	<5000	5-10000	10-30000	NO CEIL	Σ	CHANGE
<5000 ft	54%	4	2	4	64%	6%
5-10000 ft	5	4	1	1	11	2
10-30000 ft	2	1	2	1	6	3
NO CEIL	2	1	1	15	19	3
Σ	63	10	6	21	100%	14%

B) CLOUD AMOUNT (TENTH OF SKY COVER)
11 AM

8 AM GMT	0-5/10	6-9/10	10/10	Σ	CHANGE
0-5/10	15%	4	0	19	4%
6-9/10	3	11	7	21	3
10/10	2	10	48	60	2
Σ	20	25	55	100%	9%

C) VISIBILITY (Km)
11 AM

8 AM GMT	<3.2	3.2-8	3-20	>20	Σ	CHANGE
<3.2 km	20%	8	1	1	30%	10
3.2-8 km	6	12	4	1	23	6
8-20 km	1	7	14	3	25	1
>20 km	1	1	3	17	22	1
Σ	28	28	22	22	100%	18%

TABLE 2. STRUCTURE MATRIX

STUTTGART, JANUARY 1946-1953, 08 GMT

	UNROTATED				ROTATED (ORTHOGONAL)			
U	.44	.53	-.48	.49	.13	.18	-.93	.16
T	.63	.49	-.42	-.12	.36	.52	-.62	-.20
RH	.04	-.66	-.57	-.44	.07	-.32	.08	-.92
CEIL	-.89	.21	-.09	-.01	-.90	.04	.15	.06
CL AMT	.91	-.24	.04	.01	.92	-.07	-.18	-.10
Ln VIS	.10	.90	.18	-.23	-.06	.88	-.13	.32
CEIL 3	-.87	.19	-.20	.06	-.92	-.03	.04	.01
CL AMT 3	.91	-.17	.20	-.03	.94	.02	-.08	.01
Ln VIS	.14	.88	.02	-.34	-.06	.92	-.19	.14
VAR	3.81	2.70	.86	.61	3.54	2.04	1.36	1.04
VAR %	42	30	10	7	39	23	15	13

U = ZONAL WINDSPEED, T = TEMPERATURE,

RH = REL. HUMIDITY, CEIL = CEILING HEIGHT,

CL AMT = TOTAL SKY COVER,

Ln VIS = LOGARITH OF VISIBILITY

THE NUMBER 3 INDICATES THE ELEMENT 3 HOURS LATER.

TABLE 3. MEAN FACTORS BY GROUPS

CEILING HEIGHT

CEIL 8^h	CEIL 11^h	F_1	F_2	F_3	F_4	N
999	999	-8.22	.79	-.13	-.16	30
100-300	999	-4.28	-1.34	-1.44	.57	1
≤ 100	999	-2.17	.53	-.50	-.48	10
≤ 50	999	-2.40	.31	-.42	-.44	9
999	100-300	-3.34	.80	.25	-.54	4
100-300	100-300	.39	.65	.21	.81	6
≤ 100	100-300	.95	-.17	-.34	.35	7
≤ 50	100-300	1.42	.51	-.74	.40	6
999	<100	-3.36	.50	.24	-.33	5
100-300	<100	1.24	.71	.61	.30	12
<100	<100	2.10	-.31	.06	.06	125
< 50	<100	2.10	-.55	-.05	.02	115

(CEILING IN 100 ft.)

TABLE 4. GROUP SELECTION USING MEAN FACTORS

A) CEILING HEIGHT (IN 100 FT.)

CEIL 6^h	CL. AMT 6^h	CHARACT	CEIL OBS 11^h	PREDICTED Y_1	Y_4
999	0	REMAIN	999	949.6	987.9
999	1-5	CHANGE	<300	525.8	575.7
100-300	≤8	REMAIN	<300	202.2	155.4
100-300	7	CHANGE	999	607.8	609.6
<100	10	REMAIN	<100	53.1	37.0
<100	≤9	CHANGE	999	424.2	506.1

B) CLOUD AMOUNT (IN TENTH SKY COVER)

CL. AMT 6^h	CEIL 6^h	CHARACT	CL. AMT OBS 11^h	PREDICTED Y_1	Y_4
10	<30	REMAIN	10	9.6	9.8
10	30-100	CHANGE	6-9	9.2	9.2
10	>100	CHANGE	0-5	5.7	5.5
6-9	≤50	REMAIN	6-10	8.6	8.7
6-9	>50	CHANGE	0-5	6.4	6.0
4-5	999	CHANGE	6-10	4.6	6.0
0-3	999	REMAIN	0-5	1.9	1.7

Y_1 = ONE FACTOR, Y_4 = FOUR FACTORS

TABLE 5. RESIDUAL VARIANCE (IN %) FOR THREE PREDICTION MODELS

A) 200 DATA SAMPLE

	PERS	F_1	F_4	REGR	RATIO(REGR/F_1)
CEIL	55.6	22.6	16.5	43.0%	1.90
CEIL 2	54.7	23.3	16.2	41.5	1.78
VIS	50.4	23.6	20.5	69.0	2.90
CL AMT	38.9	16.3	11.4	37.9	2.32

B) 50 DATA SAMPLE (IDEAL)

CEIL	120.9	36.1	3.2	62.7%	1.74
VIS	56.6	16.9	13.4	31.4	1.86
CL AMT	89.2	34.0	8.1	67.3	1.98

C) 50 DATA SAMPLE (200 DATA COEFF.)

CEIL	126.0	39.6	35.0	86.7%	2.19
VIS	42.6	15.7	16.1	66.0	4.20
CL AMT	78.5	33.5	29.0	69.9	2.09

PERS = PERSISTENCE, F_1 = USING ONE, F_4 = USING FOUR FACTORS, REGR = REGRESSION MODEL.

FIGURE 1. STUTTGART, JANUARY 1946-1952

CEILING

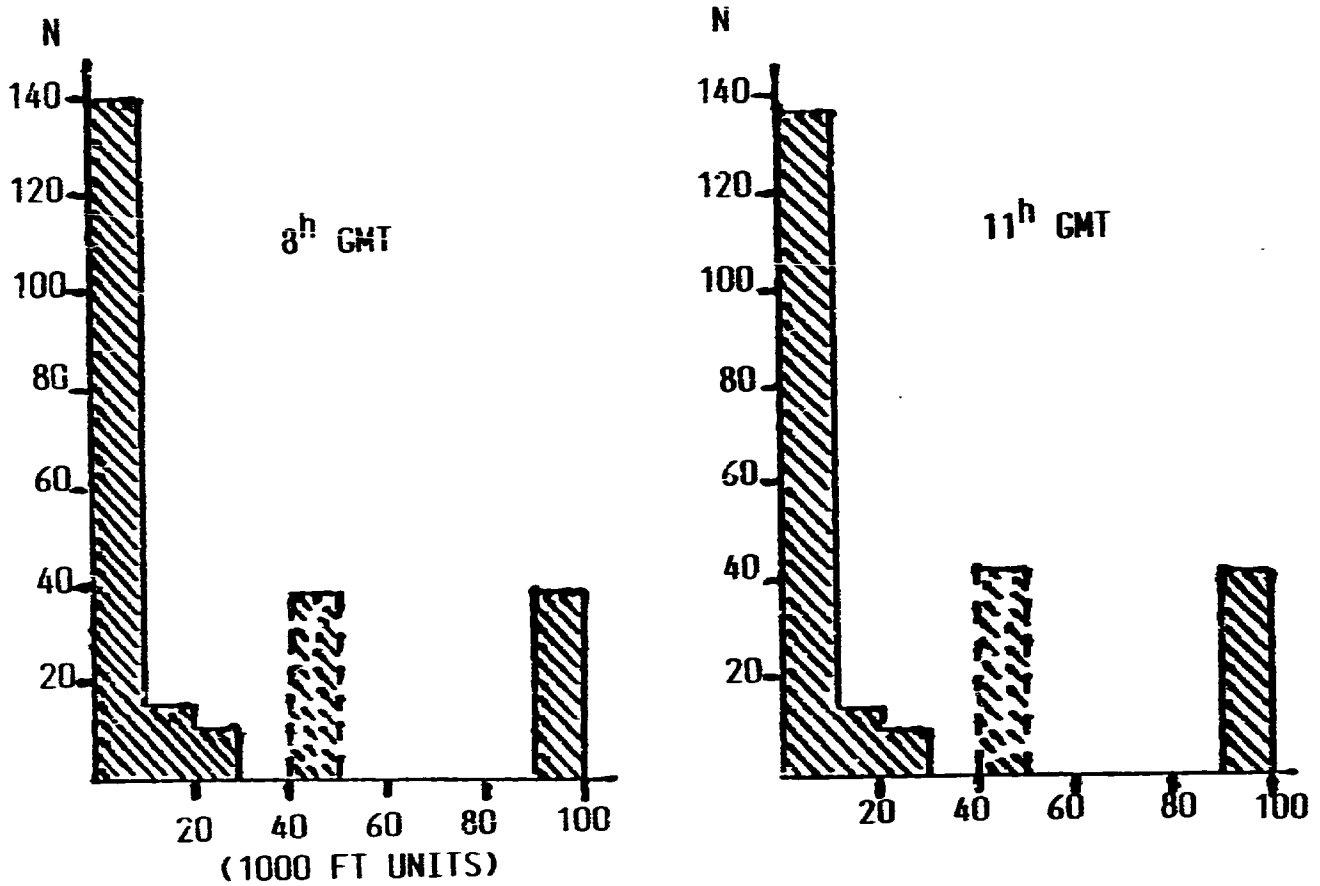


FIGURE 2. STUTTGART, JANUARY 1946-1952

SKY COVER (CLOUD AMOUNT)

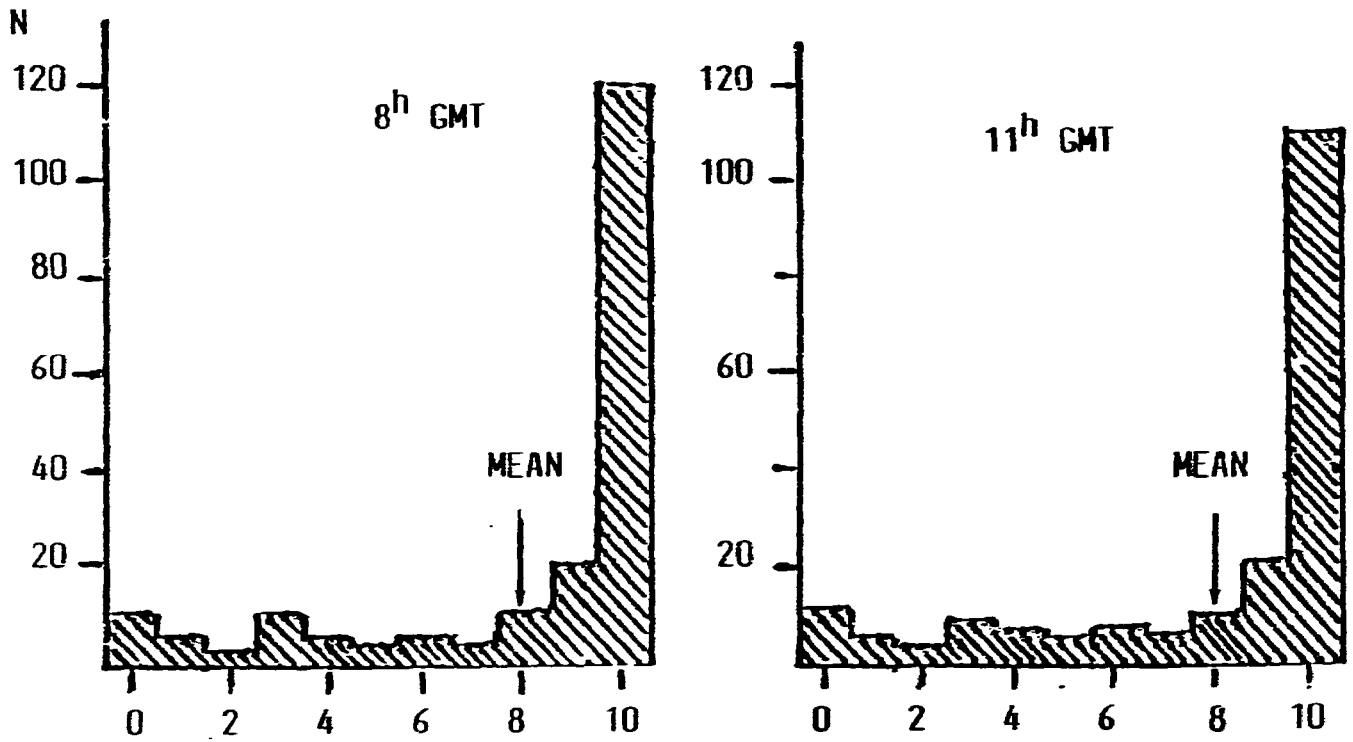
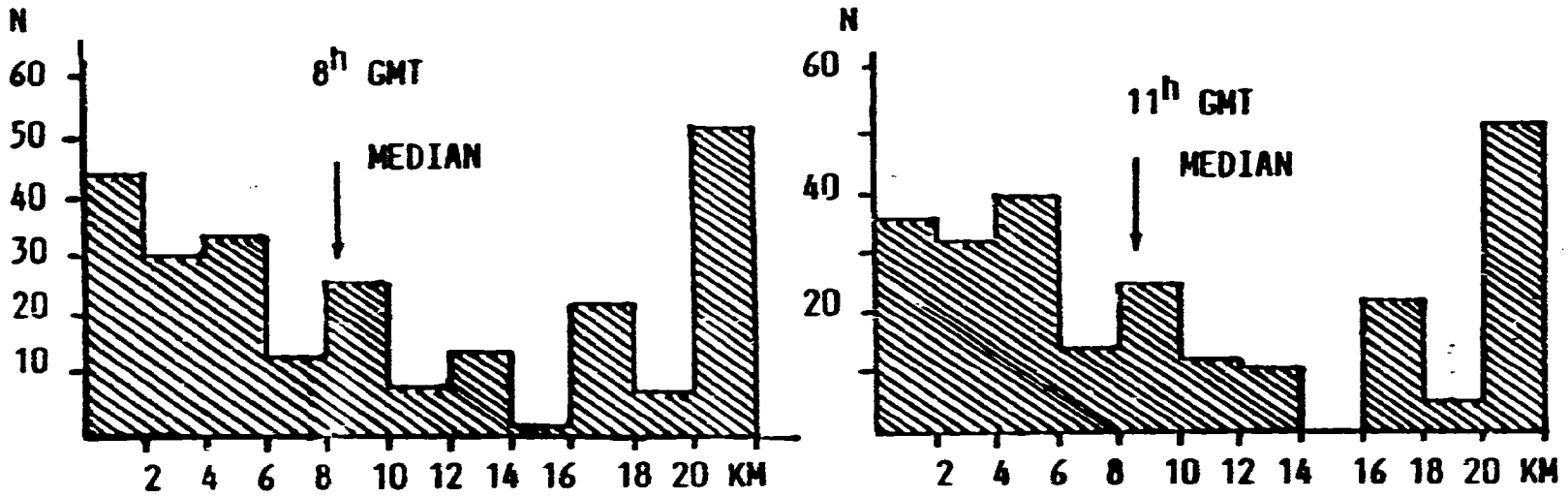


FIGURE 3. STUTTGART, JANUARY 1946-1954

VISIBILITY



CONSISTENCY OF THE P-VALUE AND A SET OF Q-VALUES IN A SCORING ACCURACY ANALYSIS

Paul H. Thrasher
U.S. Army Materiel Test and Evaluation/Engineering and Analysis RAM Division
U.S. Army White Sands Missile Range
White Sands Missile Range, New Mexico 88002-5175

ABSTRACT

One particular application, an investigation of bias in a scoring device, illustrates the use of p-value and q-value analyses. The q-values, the post-test estimates of Type II risks, are used to estimate a bias. This estimation is shown to be meaningful by the consistency of different analyses.

INTRODUCTION

Hypothesis testing is a well established analysis technique. This fairly rigid procedure can be outlined in distinct steps: [1]

(1) State a null hypothesis H_0 and an appropriate alternate hypothesis H_1 regarding a parameter θ.

(2) Specify the acceptable Type I risk α of falsely rejecting H_0, the acceptable Type II risk β of falsely failing to reject H_0 when θ has an unacceptable parameter θ_U, and the planned sample size n_p by using the sampling distribution of an appropriate test statistic.

(3) Obtain sample data.

(4) Decide and report whether or not to reject H_0.

In the traditional hypothesis testing technique, the report of this reject or not-reject decision conveys no information concerning the strength of the evidence for the decision. There are, however, two methods that can be used simultaneously to describe the evidence for rejection or non-rejection of H_0.

One method of indicating the strength of the decision is to calculate and report the p-value.[8] The p-value is the smallest value of α that would have allowed the sample data to cause H_0 to be rejected. A very low p-value strongly implies rejection of H_0.

A second method of indicating the strength of the decision is to calculate and report a q-value for θ_U.[9] The q-value is the output of the algorithm that was used to find β when the algorithm inputs α and n_p are replaced by the p-value and the data sample size. A very high q-value strongly implies rejection of H_0 in favor of H_1 characterized by θ_U.

It is possible to combine the p-value and a q-value in a single measure of evidence for rejection of H_0. One combined measure is the ratio of a q-value to the p-value.[9] A more informative combined measure is the ratio (q-value/β)/(p-value/α) or (q-value/p-value)/(β/α).[8]

For analyses in which α, β, and especially θ_U are not firmly established, the most flexible and meaningful approach is to consider the post-test Type I and Type II risks separately. Since there is a q-value for every θ_U, the analyst should report the p-value and a set of q-values corresponding to a set of θ_U's of possible interest. When these two methods are used simultaneously, a decision can be based on a comprehensive view of the evidence.

APPLICATION

The data for the application discussed in this paper is presented in Table 1. These data are estimates of Cartesian coordinates for points in a vertical plane. The abscissa is horizontal and the ordinate is vertical. Estimates are reported from both a scoring device and a standard. The scoring device is expected to have different horizontal and vertical characteristics because of physical effects. The standard is more than an order of magnitude more accurate than that which is expected of the scoring device. The two partial scores of the scoring device are not independent. Each is obtained from two intermediate results and one intermediate result is shared by the two partial scores. The final result of the scoring device is normally obtained by averaging the two partial scores. This is not done here because

(1) the drop-outs of the 25 points do not coincide so averaging would further decrease the sample size, and

(2) comparison of the results from the two partial scores can tentatively provide a check for consistency.

The primary approach used in this application is to do a p-value and q-value analysis on the parameters describing scaling and fixed biases. Linear regression is used to find least-squares estimates of A and B in $y = Ax + B$ where y is the scoring device data and x is the standard data. Separate calculations are done on both

(1) horizontal and vertical data and

(2) partial scores.

The parameter A should be unity if there is no scaling bias, and B should be zero if there is no fixed bias.

Table 2 contains results of a least-squares fit of a straight line to the data. The coefficients of correlation are sufficiently low to suggest that the fit is inadequate to specify A and B without reservations. Further indications of reservations are obtained by considering the ranges that are overlapped by the estimates of A and B plus and minus the corresponding standard deviations. All four slopes are close to one, but the slopes for vertical data have high standard deviations which overlap not only one but values quite different from one. The intercepts for horizontal data are close to zero, and the standard deviations overlap zero. The intercepts for vertical data are above zero and their standard deviations, even though they are large, do not overlap zero. The standard deviations of the means, obtained by dividing the square roots of the sample sizes into the standard deviations of data from the line, are all near or less than 0.4 meter. This implies that the random error of the scoring device is near or less than 0.4 meter.

Table 3 contains the results of one-sided, Student's-t hypothesis tests on B. All null hypotheses assume no fixed bias. The direction of each alternate hypothesis was obtained from the sign of the data average. For horizontal data from both partial scores, the p-values are sufficiently high and the q-values, for possible biases further from zero than 0.2 meter, are sufficiently low to suggest that there is no fixed bias. For vertical data, rejection for p-values less than 0.10 and q-values greater than 0.30 suggests that there may be a fixed bias of 0.6 meter to 1.2 meter. This agrees with the point estimates tentatively suggested in Table 2.

Table 4 contains the results of one-sided, Student's-t hypothesis tests on A. All null hypotheses assume no scaling bias. For both horizontal and

vertical data, the p-values are sufficiently high and q-values, corresponding to possible biases in the range of 0.8 1/m to 1.2 1/m, are sufficiently low to suggest that there is no scaling bias.

An alternate approach used in this application is to investigate the bias by doing a p-value and q-value analysis on Δ where Δ is the difference between the scoring device and standard estimates of point location. These differences are obtained by subtraction of data from Table 1.

Table 5 contains the result of one-sided, Student's-t hypothesis tests on Δ. A mean of zero would indicate no bias. For horizontal data from both partial scores, the p-values are sufficiently high and the q-values, corresponding to possible biases further from zero than 0.2 meter, are sufficiently low to suggest that there is no bias. For vertical data, rejection for p-values less than 0.10 and q-values greater than 0.30 suggests that there may be a bias of 0.6 meter to 1.2 meter. This is in agreement with the point estimates tentatively suggested in Table 2 and with the p-value and q-value analysis of Table 3.

CONSISTENCY

This example illustrates the consistency of p-value and q-value analyses. There certainly are issues that need investigation before the general technique is judged to be universally applicable and reliable. One issue is the effect of using critical levels of significance other than 0.10 and 0.30 for the post-test Type I and Type II errors. A more serious issue is the need for a comprehensive study on the properties of the q-value. This study should

include both theoretical and simulation investigations. It should consider such factors as different underlying distributions and sensitivity to extraneous data. In the absence of such a study, however, this paper provides an example of consistency in the p-value and q-value analysis technique.

Table 6 repeats information from Tables 3 and 5 in a format to allow easy comparison between the two hypothesis tests on fixed bias B and total bias A. Based on the retention of the null hypothesis that there is no scaling bias, these two tests should give the same results.

When a decision needs to be made, the q-values are in close agreement for the two hypothesis tests. For vertical data, the p-values and q-values differ only slightly for the two tests for bias.

For horizontal data, the agreement is not as good. In this case, however, rejection is not warranted. This is indicated by sufficiently high p-values and the sufficiently low q-values for biases bigger than the estimated 0.4 meter random error of the scoring device. Thus, for horizontal measurements, q-values are not needed to estimate the size of the bias.

The results of the two p-value and q-value analyses are consistant where consistancy is needed. Thus, this example supports the hypothesis that the p-value and q-value analysis is meaningful.

CONCLUSION

This application illustrates the value of the p-value and q-value analysis. This type of analysis should be done to consider and report the best post-test estimates of both Type I and Type II risks. Analysts should provide managers with this information so managers can make informed decisions.

References

[1]FREUND, John E. and WALPOLE, Ronald E., (1980) Mathematical Statistics, Prentice-Hall, Inc.

[2]IMAN, R. L. and CONOVER, W.J., (1983), A Modern Approach to Statistics, John Wiley and Sons.

[3]THRASHER, P. H., (1984) "Modification of Alpha and Beta in Hypothesis Testing," Proceedings of the Thirtieth Conference on the Design of Experiments in Army Research, Development, and Testing, U.S. Army Research Office.

[4]DUNCAN, Richard H. and THRASHER, Paul H., (1985) "Application of Hypothesis Testing to Performance Appraisal," Proceedings of the Thirty-First Conference on the Design of Experiments in Army Research, Development, and Testing, U.S. Army Research Office.

[5]THRASHER, Paul H., (1986) "Use of the P-Value and a Q-Value in Rejection Criteria," Proceedings of the Thirty-Second Conference on the Design of Experiments in Army Research, Development, and Testing, U.S. Army Research Office.

TABLE 1.--Data for scoring device calibration

Point Identification		Data from Standard (meters)		Data from Scoring Device (meters)			
				Partial Score One		Partial Score Two	
Group	Point	Horizontal	Vertical	Horizontal	Vertical	Horizontal	Vertical
1	1	1.0	1.0	1.0	1.2	1.8	1.8
2	1	0.8	0.0	1.8	-2.8	1.5	0.7
	2	1.1	-1.8	1.4	-0.4	0.7	-1.4
	3	0.0	-1.2	-	-	-0.1	-0.9
	4	1.6	-0.4	1.3	-0.7	1.1	-1.4
	5	0.3	1.0	-0.2	2.3	-0.3	0.9
	6	1.5	0.6	-	-	1.4	1.2
	7	1.0	0.4	2.3	5.0	1.0	0.5
	8	0.7	0.4	0.6	0.9	0.7	0.3
3	1	1.0	2.1	1.2	2.6	1.1	4.3
	2	0.5	-0.3	-0.5	4.2	0.1	1.2
	3	-0.5	0.4	-0.8	1.6	-0.4	0.5
	4	-0.2	0.5	0.0	0.6	-0.2	0.9
	5	-0.3	0.7	-0.2	1.4	-	-
	6	0.2	0.7	0.1	1.0	-	-
	7	-0.3	0.7	0.2	1.9	-0.1	0.7
	8	0.2	-0.5	0.1	0.5	0.0	0.7
4	1	1.0	0.9	0.0	2.8	0.8	1.1
	2	0.7	-0.4	0.3	2.2	-	-
	3	-0.4	-0.3	-0.3	-1.6	-0.3	0.7
	4	0.0	-0.6	0.1	-0.3	-0.3	-1.7
	5	-0.2	-0.1	-1.4	5.6	-0.2	0.1
	6	0.7	-0.5	0.4	-0.5	0.1	4.0
	7	0.6	0.6	-2.0	2.6	0.6	2.7
	8	1.4	-0.2	1.3	-2.3	1.1	-0.8

TABLE 2.--Summary of linear regression
(Least squares fit of y = Ax + B for y = scoring device & x = standard data)

Measurement: Partial Score:	Horizontal One	Horizontal Two	Vertical One	Vertical Two
Sample Size:	23	22	23	22
Correlation:	0.63	0.86	0.37	0.63
A (1/m):	1.039	0.927	0.996	1.149
B (m):	-0.201	-0.038	1.022	0.591
s_A (1/m):	0.279	0.123	0.552	0.320
s_B (m):	0.212	0.101	0.437	0.265
$s_{y\text{-line}}$ (m):	0.795	0.360	2.036	1.231
$s_{\text{mean y-line}}$ (m):	0.166	0.077	0.425	0.263

TABLE 3.--Summary of Student's-t hypothesis tests on B
(B = Intercept from y = Ax + B = fixed bias)

Measurement: Partial Score:	Horizontal One	Horizontal Two	Vertical One	Vertical Two
Null:	H_0: Mean = 0	H_0: Mean = 0	H_0: Mean = 0	H_0: Mean = 0
Alternate:	H_1: Mean < 0	H_1: Mean < 0	H_1: Mean > 0	H_1: Mean > 0
Sample Size:	23	22	23	22
Average (m):	-0.201	-0.038	1.022	0.591
Std Deviation of Mean (m):	0.212	0.101	0.437	0.265
P-Value:	0.177	0.355	0.015	0.019
Q-Value for				
Bias = 0.2 m:	0.036	0.015	0.96	0.92
Bias = 0.4 m:	0.005	0.0002	0.92	0.76
Bias = 0.8 m:	0.0001	Close to 0	0.69	0.22
Bias = 1.2 m:	Close to 0	Close to 0	0.34	0.016
Bias = 1.6 m:	Close to 0	Close to 0	0.10	0.0006
Bias = 2.0 m:	Close to 0	Close to 0	0.02	<0.00001
Bias signs:	-	-	+	+

TABLE 4.--Summary of Student's-t hypothesis tests on A
(A = slope from y = Ax + B = scaled bias)

Measurement: Partial Score:	Horizontal One	Horizontal Two	Vertical One	Vertical Two
Null:	H_0: Mean = 1	H_0: Mean = 1	H_0: Mean = 1	H_0: Mean = 1
Alternate:	H_1: Mean > 1	H_1: Mean < 1	H_1: Mean < 1	H_1: Mean > 1
Sample size:	23	22	23	22
Average (1/m):	1.039	0.927	0.996	1.149
Std Deviation of Mean (1/m):	0.279	0.123	0.552	0.320
P-Value:	0.445	0.107	0.497	0.324
Q-Value for				
Slope = R_1 1/m:	0.48	0.57	0.47	0.62
Slope = S_1 1/m:	0.42	0.41	0.43	0.56
Slope = T_1 1/m:	0.23	0.16	0.36	0.38
Slope = U_1 1/m:	0.057	0.04	0.30	0.14
Slope = V_1 1/m:	0.0094	0.008	0.24	0.038
Slope = W_1 1/m:	0.0012	0.001	0.19	0.0076
Slope subscript:	1	2	2	1

	R:	S:	T:	U:	V:	W:
Considered biases for subscript 1:	1.05	1.10	1.25	1.50	1.75	2.00
Considered biases for subscript 2:	0.95	0.90	0.80	0.70	0.60	0.50

TABLE 5.--Summary of Student's-t hypothesis tests on Δ
(Δ = scoring device data - standard data = total bias)

Measurement: Partial Score:	Horizontal One	Horizontal Two	Vertical One	Vertical Two
Null:	H_0: Mean = 0	H_0: Mean = 0	H_0: Mean = 0	H_0: Mean = 0
Alternate:	H_1: Mean < 0	H_1: Mean < 0	H_1: Mean > 0	H_1: Mean > 0
Sample Size:	23	22	23	22
Average ($\bar{m}$):	-0.183	-0.077	1.022	0.609
Std Deviation of Mean (m):	0.777	0.354	1.989	1.208
P-Value:	0.136	0.159	0.011	0.014
Q-Value for				
Bias = 0.2 m:	0.46	0.06	0.97	0.94
Bias = 0.4 m:	0.097	<0.00001	0.93	0.79
Bias = 0.8 m:	0.0048	Close to 0	0.70	0.23
Bias = 1.2 m:	<0.00001	Close to 0	0.34	0.016
Bias = 1.6 m:	Close to 0	Close to 0	0.089	0.0005
Bias = 2.0 m:	Close to 0	Close to 0	0.014	<0.00001
Bias signs:	-	-	+	+

TABLE 6.--Consistency of P-values and Q-values
(Comparison of results from hypothesis tests on B and Δ)

Measurement: Partial Score:	Horizontal One	Horizontal Two	Vertical One	Vertical Two
P-Value for B:	0.177	0.355	0.015	0.019
P-Value for Δ:	0.136	0.159	0.011	0.014
Q-Value for				
Bias = 0.2 m				
for B:	0.036	0.015	0.96	0.92
for Δ:	0.46	0.06	0.97	0.94
Bias = 0.4 m				
for B:	0.005	0.0002	0.92	0.76
for Δ:	0.097	<0.00001	0.93	0.79
Bias = 0.8 m				
for B:	0.0001	Close to 0	0.69	0.22
for Δ:	0.0048	Close to 0	0.70	0.23
Bias = 1.2 m				
for B:	Close to 0	Close to 0	0.34	0.016
for Δ:	<0.00001	Close to 0	0.34	0.016
Bias = 1.6 m				
for B:	Close to 0	Close to 0	0.10	0.0006
for Δ:	Close to 0	Close to 0	0.089	0.0005
Bias = 2.0 m				
for B:	Close to 0	Close to 0	0.02	<0.00001
for Δ:	Close to 0	Close to 0	0.014	<0.00001

DENSITY ESTIMATION, MODELING AND SIMULATION: STUDIES IN EMPIRICAL MODEL BUILDING*

Lectures presented at the Thirty-second Conference on the Design of Experiments in Army Research, Design and Testing, Monterey, California, October 27-28, 1986.

James R. Thompson, Rice University, Houston, Texas 77251-1892

*This work was supported in part by the Army Research Office (Durham) under DAAG-29-85-K-0212 at Rice University

The editors of these Proceedings would like to thank Professor James R. Thompson for his permission to reproduce here the following sections of this series of lectures: Chapter 2, Sections 3 through 6, and all of Chapter 4.

Dedication: to Major Henryk Sucharski
Virtuti Militari (1920) Polish-Soviet War; Officer Commanding, Westerplatte Garrison, September 1-7, 1939, Polish-German War; on the Fortieth Anniversary of His Death.

Być pobitym, ale się nie poddać--ale zwycięstwo. Józef Piłsudski.

UBI HABITAT LIBERTAS, IBI NOSTRA PATRIA EST. Motto on the Texian Flag at the Battle of San Jacinto.

Acknowledgements: Neely Atkinson, Barry Brown, Steve Boswell, Andrew Donoho, Ed Johnson, Frank Jones, Tom Kauffman, Robert Launer, David Scott, Richard Tapia, Malcolm Taylor, George Terrell, John Tukey.

Table of Contents

PREFACE

The study of mathematical models is closely connected to notions of scientific creativity. As of the present, there is no axiomatic or even well defined discipline which is directly concerned with creativity. Even though we cannot display a progression of exercises which have as their direct objective the building of creativity, we can attempt to accomplish this goal indirectly. A mastery of a portion of Euclid's treatises on geometry does not directly appear to build up a potential statesman's ability to practise statecraft. Yet many effective statesmen have claimed that their studies of Euclid's geometry had achieved this effect. More directly, it is clear that the study of physics would be likely to be helpful in developing the ability to design good automobiles. It is this carryover effect from one well defined discipline to another less defined one which has traditionally been the background of science and engineering education.

Valuable though an indirect approach to the gaining of creativity in a particular area may be, it carries with it certain dangers. We are rather in the same situation as the little boy who searched for his quarter, lost in a dark alley, under a bright streetlight on a main street. There is no doubt that the main-street searching could be of great utility in the ultimate quest of finding the quarter. Many of the relevant techniques in quarter finding are similar, whether one is looking in the light or in the dark. Hopefully, the study of technique, albeit undertaken in a setting substantially different from that of the real problem, will be at least marginally useful in solving the real problem.

However, there is a natural temptation never to leave the comfort of idealized technique under the bright lights, never to venture into the murky depths of the alley where the real problem lies. How much easier to stay on mainstreet, to write a treatise on the topology of street lamps, gradually to forget about the lost quarter altogether.

In its most applied aspect, technique becomes problem solving. For example, if the little boy really develops a procedure for finding his particular quarter in the particular dark alley where he lost it, he will have been engaged in problem solving. Although it is difficult to say where problem posing ends and problem solving begins, since in the ideal state there is continuous interaction between the two, model building is more concerned with the former than with the latter. Whereas problem solving can generally be approached by more or less well defined techniques, there is seldom such order in the problem posing mode. In the quarter finding example, problem posing would involve determining that it was important that the quarter be found and a description of the relevant factors concerning this task. Here, the problem posing is heuristic, difficult to put into symbols and trivial. In the real world of science, problem posing is seldom trivial, but remains generally heuristic and difficult to put into symbols. For example, Newton's Second Principle states that force is equal to the rate of change of momentum or

$$F = \frac{d}{dt}(mv). \tag{0.1}$$

The solving of (0.1) for a variety of scenarios is something well defined and easily taught to a high school student. But the thought process by which Newton conjectured (0.1) is far more complex.

We have no philosopher's stone to unlock for us the thought processes of the creative giants of science. And we shall not use the device of scientific biography much in this treatise. However, the case study approach appears to be useful in the development of creativity. By processes which we do not understand, the mind is able to synthesize the ability to deal with situations apparently unrelated to any of the case studies considered. It is the case study approach, historically motivated on occasion, which we shall emphasize.

At this point, it is appropriate that some attempt be made to indicate what the author means by the term *Empirical Model Building* . To do so, it is necessary that we give some thought to some of the ways various scientists approach the concept of models. We shall list here only those three schools which appear to have the greatest numbers of adherents. The first group we shall term the *Idealists*. The Idealists are not really data oriented. They are rather concerned with theory as a mental process which takes a cavalier attitude toward the "real world." Their attitude can be summed up by, "If facts do not conform to theory, then so much the worse for facts." For them, the "model" is all. An example of a pure Idealist is given by the character of Marat in Weiss' play *MaratSade*. Marat says "Against Nature's silence I use action. In the vast indifference I invent a meaning." Although Idealists do crop up from time to time in the physical and biological sciences, they have a hard time there. Sooner or later, the theories of a

Lysenko, say, are brought up against the discipline imposed by the real world and must surrender in the face of conflicting evidence. But even in the hard sciences, there is the possibility that "sooner or later" may mean decades. Once a theory has developed a constituency of individuals who have a vested interest in its perpetuation, particularly if the theory has no immediate practical implication, there will be a tendency of other scientists, who have no interest in the theory either way, to let well enough alone.

The second group, that of the *Radical Pragmatists* (Occamites, nominalists) would appear to be at the opposite end of the spectrum from that of the Idealists. The Radical Pragmatists hold that data is all. Every situation is to be treated more or less *sui generis*. There is no "truth." All models are false. Instead of model building, the Radical Pragmatist curve fits. He does not look on his fitted curve as something of general applicability, rather as an empirical device for coping with a particular situation. The maxim of William of Occam was "Essentia non sunt multiplicanta praeter necessitatem," roughly, "The hypotheses ought not to be more than is necessary." The question here is what we mean by "necessary." All too frequently, it can happen than "necessary" means what we need to muddle through rather than what we need to understand. But few Radical Pragmatists would take the pure position of Weiss's Sade who says "No sooner have I discovered something than I begin to doubt it and I have to destroy it again...the only truths we can point to are the ever-changing truths of our own experience."

The *Realists* (Aristoteleans, Thomists) might appear to some to occupy a ground intermediate to that of the Idealists and that of the Radical Pragmatists. They hold that the universe is governed by rational and consistent laws. Models, for the Realist, are approximations to bits and pieces of these laws. To the Realist, "We see through a glass darkly," but there is reality on the other side of the glass. The Realist knows his model is not quite right, but he hopes it is incomplete rather than false. The collection of data is useful in testing his model and enabling him to modify it in appropriate fashion. It is this truthseeking, interactive procedure between mind and data which we term *Empirical Model Building* .

To return again to Newton's Second Principle, the position of the Idealist might be simply that the old Newtonian formula

$$F = ma \tag{0.2}$$

is true because of logical argument. But then we have the empirically demonstrable discovery of Einstein that mass is not constant but depends on velocity via

$$m = \frac{m_0}{\left[1-\frac{v^2}{c^2}\right]^{\frac{1}{2}}} v. \tag{0.3}$$

The Idealist would have a problem. He might simply stick with (0.2) or experience an intellectual conversion, saying, "Right, Einstein is correct; Newton is wrong. I am no longer a Newtonian but an Einsteinian" (or some less self-effacing dialectical version of the above conversion.)

The reaction of the Radical Pragmatist might be, "You see, even an apparently well established model like Newton's is false. No doubt we will soon

learn that Einstein's is false also. Both these 'models' are useful in many applications, but their utility lies solely in their applicability in each situation."

The Realist is also unsurprised that Newton's model falls short of the mark. He notes that the discovery of Einstein will require a modification of (0.2). He readily accomplishes this by combining (0.1) and (0.3) to give

$$F = \frac{d}{dt}\frac{m_0}{[1-\frac{v^2}{c^2}]^{\frac{1}{2}}}v \tag{0.4}$$

He views (0.4) as a better approximation to truth than (0.2) and expects to hear of still better approximations in the future.

The preceeding should give the reader some feel as to what the author means by empirical model building (and also as to his prejudices in favor of the Realist position). It is the process which is sometimes loosely referred to as the "scientific method." As such, it has been around for millenia--though only for the last five hundred years or so has quantitative data collecting enabled its ready use on nontrivial scientific problems. Realists might argue (as I do) that empirical model building is a natural activity of the human mind. It is the interactive procedure by which human beings proceed to understand portions of the real world by proposing theoretical mechanisms, testing these against observation and revising theory when it does not conform to data. In any given situation, a scientist's empirical model is simply his current best guess as to the underlying mechanism at hand.

The Radical Pragmatist position has great appeal for many, particularly in the United States. There would appear to be many advantages to an orientation which allowed one to change his ground any time it was convenient to do so. But the ultimately nihilistic position of Radical Pragmatism has many practical difficulties. For example, data is generally collected in the light of some model. Moreover, from the standpoint of compression of information, a point of view which rejects truth also rejects uniqueness, causing no little chaos in representation. Finally, the old adage that "He who believes in nothing will believe anything" appears to hold. The Radical Pragmatist seems to join hands with the Idealist more often than either cares to admit. There are certain groups who seem to wear the colours of both the Idealist and Radical Pragmatist schools.

The above taxonomy of contemporary scientists into three fairly well defined schools of thought is, obviously, an oversimplification. Most scientists will tend to embody elements of all of the three schools in their makeup. For example, I might be (and have been) accosted in my office by someone who wishes me to examine his plans for a perpetual motion machine or his discovery of a conspiracy of Freemasons to take over the world. As a purely practical matter, because my time is limited, I will be likely to dismiss their theories as patently absurd. In so doing, I am apparently taking an Idealist position, for, indeed I know little about Freemasonry or about perpetual motion machines. But without such practical use of prejudice, nothing could ever be accomplished. We would spend our lives "starting from zero" and continually reinventing the wheel. There is a vast body of information which I have not investigated and yet take to be true, without

ever carefully checking it out. This is not really "Idealism"; this is coping. But if I read in the paper that Professor Strepticollicus had indeed demonstrated a working model of a perpetual motion machine, or if I heard that a secret meeting room, covered with Masonic symbols, were discovered in the Capitol, then I should be willing to reopen this portion of my "information bank" for possible modification.

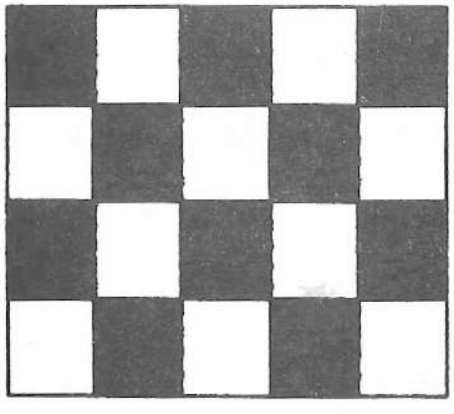
For similar practical reasons, I must act like a Radical Pragmatist more often than I might wish. If I see a ten ton truck bearing down on me, I will instinctively try to get away without carefully investigating considerations of momentum and the likely destruction to human tissue as a result of the dissipation thereof. But I have the hope that the manufacturer of the truck has logically and with the best Newtonian theory in tandem with empirical evidence designed the vehicle and not simply thrown components together, hoping to muddle through.

In sum, most of us, while accepting the practical necessity of frequently assuming theories which we have not analyzed and using a great deal of instinctive rather than logical tools in our work, would claim to believe in objective reality and a system of natural laws which we are in a continuing process of perceiving. Thus, most of us would consider ourselves to be Rationalists though we might, from time to time, act otherwise. Perhaps the minimal Rationalist maxim is that of Orwell's Winston Smith "Freedom is the freedom to say that two plus two make four. If that is granted, all else follows."

Section 3. Modular Wargaming

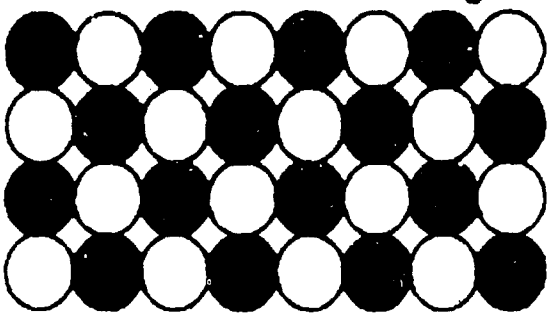
Checkerboard based games are of ancient origin, being claimed by a number of ancient cultures. One characteristic of these games is the restricted motion of the pieces, due to the shape of the playing field. This is overcome, in measure, in chess, by giving pieces varying capabilities for motion both in direction and distance. Another characteristic of these games is their essential equality of firepower. A pawn has the same power to capture a queen as the queen to capture a pawn. Effectiveness of the various pieces is completely a function of their mobility.

Figure 1



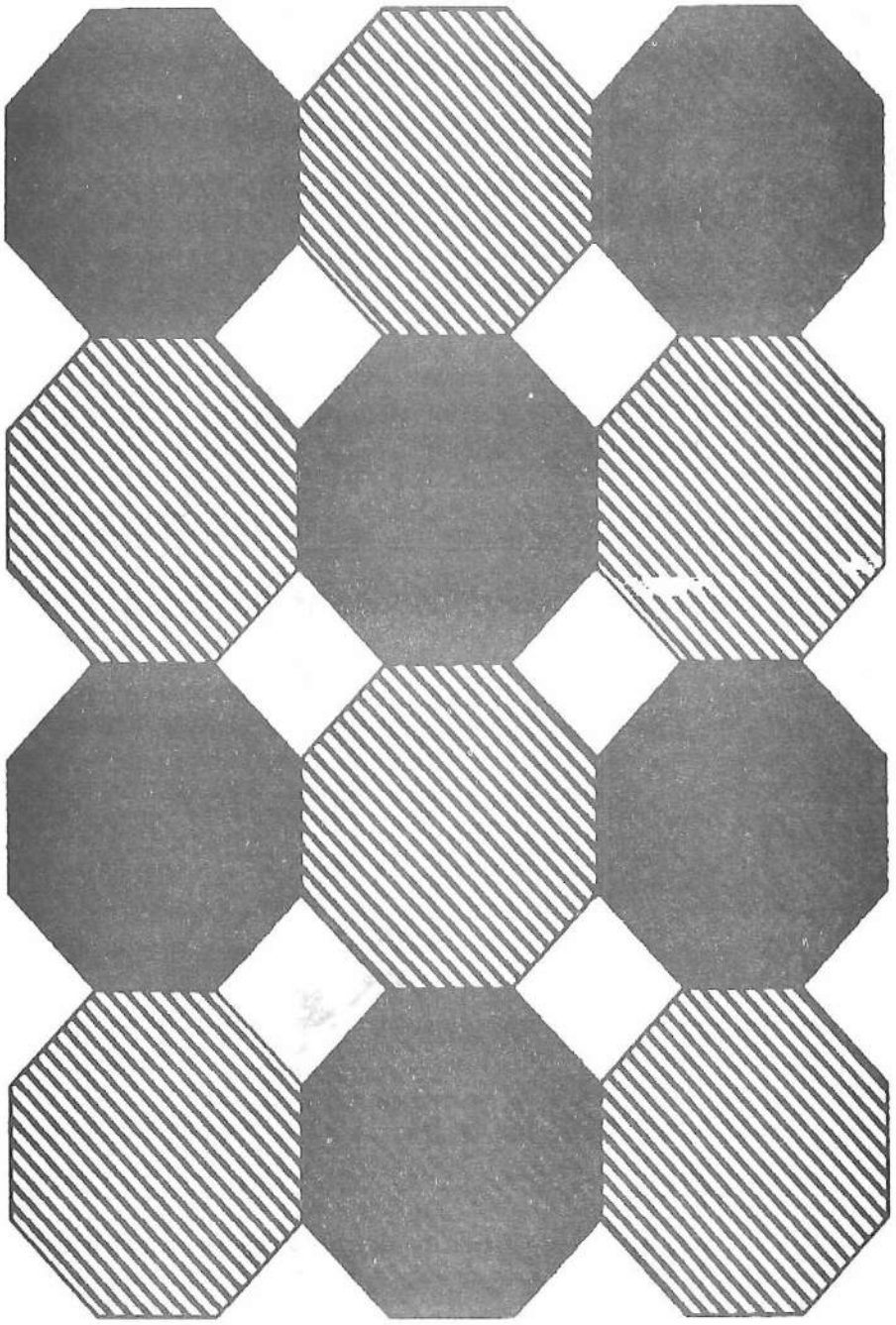
The directional restrictions of square tiles are a serious detriment to checkboard games if they are to be reasonable simulations of warfare. The most satisfactory solution, at first glance, would appear to be to use building blocks based on circles, since such tiles would appear to allow full 360 degree mobility. Unfortunately, as we observe below, circles cannot be satisfactory tiles, since they leave empty spaces between the tiles.

Figure 2



A natural first attempt to overcome the difficulty of circles as tiles would be to use equilateral octagons, since these allow motion to the eight points of the compass, N,NE,E,SE,S,SW,W,NW. Unfortunately, as we see below, this still leaves us with the empty space phenomenon.

Figure 3



None of the ancient games is particularly apt as an analogue of combat after the development of the longbow, let alone after the invention of gunpowder. Accordingly, the Prussian von Reiswitz began to make suitable modifications leading in 1820 to *Kriegspiel*. The variants of the Prussian game took to superimposing an hexagonal grid over a map of actual terrain.

Motion of various units was regulated by their capabilities in their particular terrain situation. The old notion of "turns" was retained, but at each turn, a player could move a number of units subject to a restriction on total move credits. Combat could be instituted by rules based on adjacency of opposing forces. The result of the combat was regulated by the total firepower of the units involved on both sides in the particular terrain situation. A roll of the dice, followed by lookup in a combat table gave the casualty figures together with advance and retreat information.

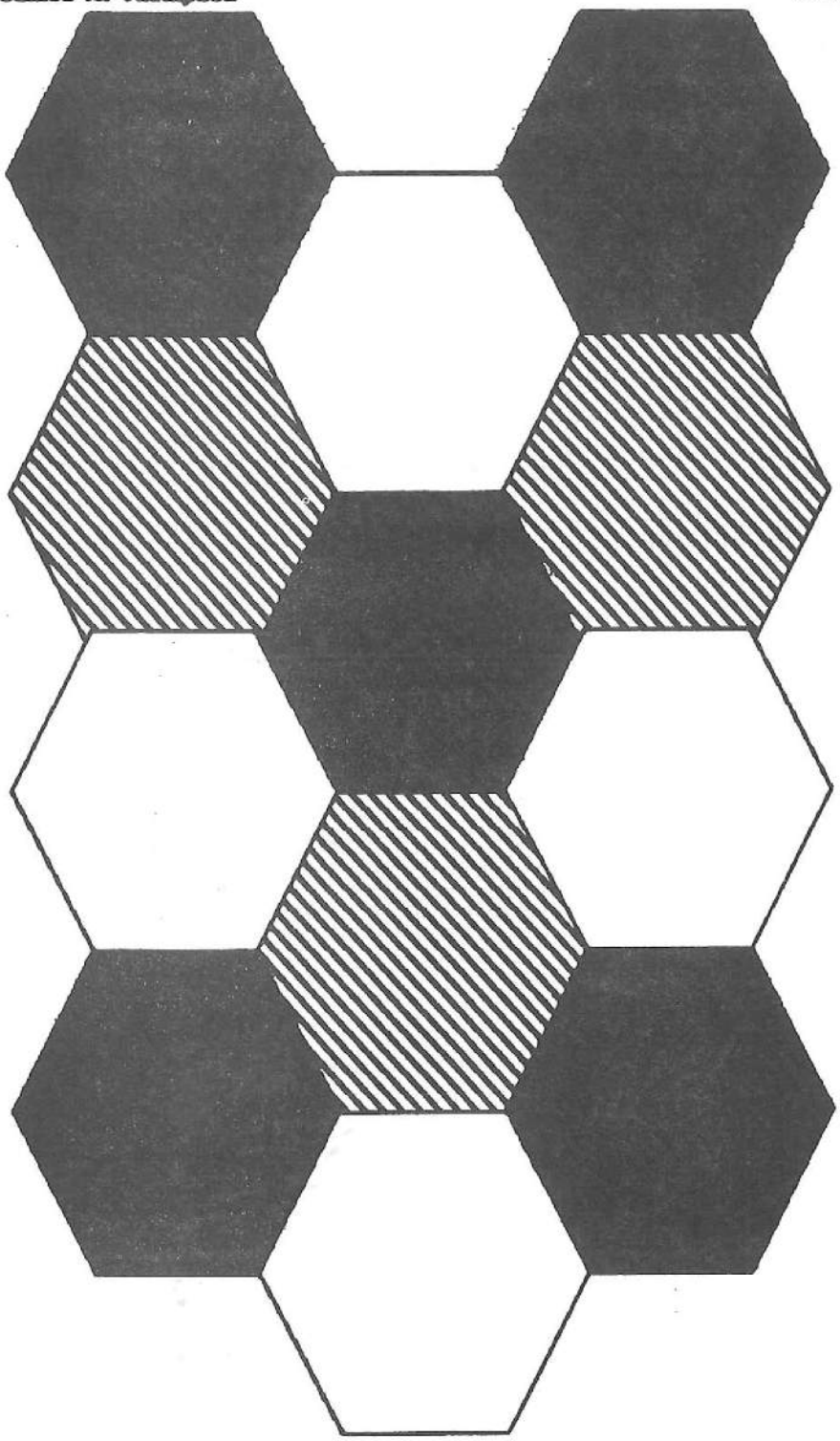


Figure 4

The Prussian game, together with later American variants, such as Strategos, were validated against actual historical combat situations. In general, these games were excellent in

their ability to simulate the real world situation. Their major difficulty was one of bookkeeping. Frequently, a simulated combat could take longer to play than the actual historical battle. If the masking of movements and questions of intelligence gathering were included in the game, a large number of referees was required.

In attempting to take advantage of the computer, the creators of many modern military wargames have attempted to go far beyond resolution of the bookkeeping problems associated with *Kriegspiel*. Very frequently, these games do not allow for any interaction of human participants at all.

Initial conditions are loaded into a powerful mainframe computer, and the machine plays out the game to conclusion based upon a complex program which may actually look at the pooled result of simulations of individual soldiers firing at each other, even though the combat is for very large units. Any real time corrections for imperfections in the game are, accordingly, impossible. Any training potential of such games is, obviously, slight.

Furthermore, the creators of many of these games may disdain to engage in any validation based on historical combat results. Such validation as exists may be limited to checking with previous generations of the same game to see whether both gave the same answer.

If we know anything about artificial intelligence (and admittedly, we know very little), it would appear to be that those simulations work best which appear to mimic the noncomputerized human system.

Attempts to make great leaps forward without evolution from noncomputerized system are almost always unsuccessful. And it

is another characteristic of such a nonevolutionary approach that it becomes quickly difficult to check the results against realistic benchmarks. Before anyone realizes it, a new, expensive, and, very likely, sterile science will have been created soaking up time and treasure and diverting us from the real world situation.

My own view is that it is better to use the computer as a means of alleviating the bookkeeping difficulties associated with *Kriegspiel*-like board games. In the late 1970's and early 1980's, I assigned this task to various groups of students at Rice. Experience showed that two hundred person hours of work generally led to games which could emulate historical results very well.

At least another five hundred person hours would have been required to make these games user-friendly, but the rough versions of the games were instructive enough. One criticism made against historical validation is that technology is advancing so rapidly that any such validations are meaningless.It is claimed that the principal function of wargaming ought to be predictions of what will happen given the new technologies. While not agreeing that parallels between historical situations and future conflicts are irrelevant (and I note here that the *Strategy and Tactics* hobbyists generally make games ranging from Bronze Age warfare to *Starship Troopers*), I agree that the predictive aspect, in the form of scenario analyses, is very important.

Accordingly, one student created a game for conflict between an American carrier task force and a Soviet missle cruiser task force. Given the close-in combat which would be likely, it appeared that if the Soviet commander is willing to sacrifice his

force for the much more costly American force, he can effect an exchange of units by a massive launch of missles at the outset of the conflict. Clearly, such a playout could have serious technological implications, e.g., the desirability of constructing a system of jamming and antimissle defenses which is highly resistant to being overwhelmed by a massive strike. Or, if it is deemed that such a system could always be penetrated by futher technological advances on the Soviet side, it might be appropriate to reconsider task forces based around the aircraft carrier. In any event, I personally would much prefer an interactive game in which I could see the step by step results of the simulation.

Also, a validation using, say, data from the Falkland conflict could be used to check modular portions of the game. World War II data could be used to check other parts. The validation would not be as thorough as one might wish, but it would be a goodly improvement on no validation at all. Some "supersophisticated" unvalidated computer simulation in which the computer simply played with itself and, at the end of the day, told me that existing antimissle defenses were sufficient would leave me neither comforted nor confident.

An integral part of any *Kriegspiel* computerization should deal with the resolution of the likely results of a conflict. A ready means of carrying this out was made available via the famous World War I opus of Lanchester (1916). Let us suppose that there are two forces, the Blue and the Red, each homogeneous, and with [redacted] sizes u and v respectively.

Then, if the fire of the Red force is directed, the probability a particular Red combatant will eliminate some Blue combatant in time interval $[t,t+\Delta]$ is given simply by:

(2.3.1) P[Blue combatant eliminated in $[t,t+\Delta]$= $c_1\Delta$,

where c_1 is the Red coefficient of undirected fire. If we wish, then, to obtain the total number of Blue combatants eliminated by the entire Red side in $[t,t+\Delta]$, we will simply multiply by the number of Red combatants to obtain:

(2.3.2) E[Change in Blue in $[t,t+\Delta]$] = $-v\, c_1\Delta$.

Replacing u by its expectation (as we have the right to do in many cases where the coefficient is truly a constant and v and u are large), we have:

(2.3.3) $\Delta u/\Delta = - c_1 v$.

This gives us immediately the differential equation

(2.3.4) $du/dt = - c_1 v$.

Similarly, we have for the Red side

(2.3.5) $dv/dt = - c_2 u$.

This system has the time solution

(2.3.6) $u(t)=u_0\cosh\sqrt{(c_1c_2)}t-v_0\sqrt{(c_1/c_2)}\sinh\sqrt{(c_1c_2)}t$

$$v(t)=v_0\cosh\sqrt{(c_1c_2)}t-u_0\sqrt{(c_2/c_1)}\sinh\sqrt{(c_1c_2)}t$$

A more common representation of the solution is obtained by dividing (2.3.4) by (2.3.5) to obtain

(2.3.7) $du/dv = c_1 v/ c_2 u$,

with the solution

(2.3.8) $u^2 - u_0^2 = c_1/c_2(v^2 - v_0^2)$.

Now u and v are at "combat parity" with each other when

(2.3.9) $u^2 = c_1/c_2 (v^2)$.

(A special point needs to be made here. Such parity models assume that both sides are willing to bear the same proportion of losses. If such is not the case, then an otherwise less

numerous and less effective force can still emerge victorious. For example, suppose that the Blue force versus Red force coefficient is .5 and the Blue force has only .9 the numerosity of the Red force. Then if Blue is willing to fight until reduced to .5 of his original strength, but Red will fight only to .8 of his original strength, then using (2.38) that by the time Red has reached maximal acceptable losses, Blue still has 61% of his forces, and thus wins the conflict. This advantage to one force to accept higher attrition than his opponent is frequently overlooked in wargame analysis. The empirical realization of this fact has not escaped the attention of guerilla leaders from the Maccabees to the Mujaheddin.)

Accordingly, it is interesting to note that if there is a doubling of numbers on the Red side, Blue can only maintain parity by seeing to it that c_2/c_1 is quadrupled, a seemingly impossible task.

Lanchester's formula for undirected fire follows from similar Poissonian arguments. The probability that a Red combatant will eliminate some Blue combatant in $[t,t+\Delta]$ is given by

(2.3.10) P[a Blue eliminated by a Red in $[t,t+\Delta]$] =

P[shot fired in $[t,t+\Delta]$] P[shot hits a Blue] Δ.

Now, the probability a shot aimed at an area rather than an individual hits someone is proportional to the density of Blue combatants in the area, hence proportional to u. Thus, we have:

(2.3.11) P[Blue eliminated in $[t,t+\Delta]$] = $d_1 u \Delta$.

The expected number of Blues eliminated in the interval is given by multiplying the above by the size of the Red force,

namely, v.

So the differential equations are:

(2.3.12) $du/dt = -d_1uv$

$dv/dt = -d_2uv.$

This system has the time solution:

(2.3.13) $u(t) = [d_2/d_1\ u_0 - v_0]/[d_2/d_1 - v_0/u_0 \exp[-(d_2\ u_0 - d_1\ v_0)t]]$

$v(t) = [d_1/d_2 v_0 - u_0]/[d_1/d_2 - u_0/v_0 \exp[-(d_1\ v_0 - d_2 u_0)t]].$

Here, when dividing the equations in (2.3.12) and solving, we obtain the parity equations:

(2.3.14) $u - u_0 = d_1/d_2(\ v - v_0\ ).$

In such a case, a doubling of Red's parity force can be matched by Blue's doubling of d_2/d_1.

In attempting to match either law (or some other) against historical data, one needs to be a bit careful. In 1954, Engel claimed to have validated the applicability of Lanchester's directed fire law for the Battle of Iwo Jima. He used no records for Japanese casualties and simply juggled the two parameters to fit the record of American casualty data.

In a STAG report written in 1972 (later published in the open literature in 1979), Thompson, using the partial Japanese casualty records, showed that the directed fire model gave answers much at variance with the data (sometimes off the Japanese total effectives by a factor of four) and that the undirected fire model appeared to work much more satisfactorily. However, the bottom line in the Thompson paper was that a homogeneous force model was probably not very

satisfactory in an engagement in which naval gunfire together with Marine assault both played important roles. We shall address the hetereogeneous force model problem directly.

In this, the one hundred and fiftieth anniversary of the Battle of the Alamo, it is perhaps instructive to consider a situation in which a mixture of the two models is appropriate. Since the Texians were aiming at a multiplicity of Mexican targets and using rifles capable of accuracy at long range (300m), it might be appropriate to use the directed fire model for Mexican casualties. Since the Mexicans were using less accurate muskets (100m) and firing against a fortified enemy, it might be appropriate to use the undirected fire model for Texian casulaties. This would give

(2.3.15) $du/dt = -d_1uv$

$dv/dt = -c_2u.$

The parity equations are given by

(2.3.16) $v^2-v_0^2 = 2c_2/d_1(u-u_0).$

The Texians fought 188 men, all of whom perished in the defense. The Mexicans fought 3,000 men of whom 1,500 perished in the attack. By plugging in initial and final strength conditions, it is an easy matter to compute $c_2/d_1 = 17,952.$ However, such an index is essentially meaningless, since the equations of combat are dramatically different for the two sides. A fair measure of man for man Texian versus Mexican effectiveness is given by

(2.3.17) $[\{dv/dt\}/u] / [\{du/dt\}/v] = c_2/(d_1u)$.

This index computes the rate of destruction of Mexicans per Texian divided by the rate of destruction of Texian per Mexican.

We note that the mixed law model gives a varying rate of effectiveness, depending on the number of Mexicans present. At the beginning of the conflict, the effectiveness ratio is a possible 96 ; at the end, a romantic but unrealistic 17,952.

The examination of this model in the light of historical data should cause us to question it. What is wrong? Most of the Mexican casualties occurred before the walls were breached. Most of the Texian casualties occurred after the walls were breached. *But after the walls were breached, the Mexicans would be using directed fire against the Texians.*

We have no precise data to verify such an assumption, but for the sake of argument, let us assume that the Texians had 100 men when the walls were breached; the Mexicans 1800. Then (2.3.16) gives $c_2/d_1 = 32{,}727$. The combat effectiveness ratio $c_2/(d_1 u)$ goes then from 174 at the beginning of the siege to 327 at the time the walls were breached. For the balance of the conflict we must use equations (2.3.4) and (2.3.5) with the combat effectiveness ratio $c_2/c_1 = 99$ (computed from (2.3.8). Personally, I am not uncomfortable with these figures. The defenses seem to have given the Texians a marginal advantage of around 3. Those who consider the figures too "John Wayneish" should remember that the Mexicans had great difficulty in focusing their forces against the Alamo, whereas the Texians were essentially all gainfully employed in the business of fighting. This advantage to a group of determined Palikari to defend a fortified position against overwhelming numbers of a besieging enemy is something we shall return to shortly.

Having, hopefully, transmitted some feeling as to the advantages of common sense utilization of the method of

Lanchester (borrowed in spirit from Poisson), we shall now take the next step in its explication: namely the utilization of heterogeneous force equations.

Let us suppose that the Blue side has m subforces $\{u_j\}_{j=1,2,\ldots,m}$. These might represent, artillery, infantry, armour, etc. Also, let us suppose that the Red side has n subforces $\{v_j\}_{j=1,2,\ldots,n}$. Then the directed fire equations (2.3.4) and (2.3.5) become:

(2.3.18) $du_j/dt = -\Sigma_{i=1 \text{ to } n} k_{ij} c_{1ij} v_i$

(2.3.19) $dv_i/dt = -\Sigma_{j=1 \text{ to } m} l_{ji} c_{2ji} u_j$.

Here, k_{ij} represents the allocation (a number between 0 and 1 such that $\Sigma_{j=1 \text{ to } m} k_{ij} \leq 1$) of the i'th Red subforce's firepower against the j'th Blue subforce. c_{ij} represents the Lanchester attrition coefficient of the i'th Red subforce against the j'th Blue subforce. Similar obvious definitions hold for $\{l_{ji}\}$ and $\{c_{2ji}\}$.

(2.3.18) furnishes us a useful alternative to the old table lookup in *Kriegspiel* Numerical integration enables us to deal handily and easily with any difficulties associated with turn to turn changes in allocation and effectiveness, reinforcements, etc. Experience has shown that computerized utilization of mobility rules based on hexagonal tiling superimposed on actual terrain, together with the use of Lanchester hetereogeneous force combat equations, makes possible the construction of realistic war games at modest cost.

Beyond the very real utility of the Lanchester combat laws to describe the combat mode for war games, they can be used as a

model framework to gain insights as to the wisdom or lack thereof of proposed changes in defense policy. In 1972 I wrote a STAG report (published in the open literature in 1979) to address the problems of disparity of NATO and Warsaw Pact forces. As we have observed in (2.3.9), in the face of a twofold manpower increase of Red beyond the parity level, Blue can, assuming Lanchester's directed fire model, maintain parity only by quadrupling c_2/c_1. This has usually been perceived to imply that NATO must rely on its superior technology to match the Soviet threat by keeping c_2 always much bigger than c_1.

Since there exists evidence to suggest that such technological superiority does not exist at the conventional level, it appears that the Soviets keep out of Western Europe because of a fear that a conventional juggernaut across Western Europe would be met by a tactical nuclear response. Thus, the big push by the Soviets and their surrogates for "non first use of nuclear weapons" treaties. It is not at all unlikely that the Soviets could take Western Europe in a conventional war.

In my paper "An Argument for Fortified Defense in Western Europe," I attempted to show how the c_2/c_1 ratio could be increased by using fortifications to decrease c_1. Whether or not the reader judges such a strategy to be patently absurd, it is instructive to go through the argument as a means of explicating the power of Lanchester's laws in scenario analysis

My investigation was motivated by the defense of the Westerplatte peninsula in Dantzig by 188 Polish soldiers from September 1 through September 7 in 1939, and some interesting parallels with the much lower tech sige of the Alamo a hundred years earlier. (Coincidentally, the number of Polish defenders

was the same as the number of Texians at the Alamo.) The attacking German forces included a batallion of SS, a batallion of engineers, a company of marines, a construction batallion, a company of coastal troops, assorted police units, 25 Stukas, the artillery of the Battleship Schleswig-Holstein, eight 150 mm howitzers, four 210 mm heavy mortars, a hundred machine guns, and two trainloads of gasoline (the Germans tried to flood the bunkers with burning gasoline).

The total number of German troops engaged in combat during the seven day seige was well over 3,000. Anyone who has visited Westerplatte (as I have) is amazed with the lack of natural defenses. It looks like a nice place for a walkover. It was not.

The garrison was defended on the first day by a steel fence (which the Germans and the League of Nations had allowed, accepting the excuse of the Polish commander, Major Sucharski, that the fence was necessary to keep the livestock of the garrison from wandering into Dantzig), which was quickly obliterated. Mainly, however, the structural defences consisted in concrete fortifications constructed at the ground level and below. Theoretically, the structural fortifications did not exist, since they were prohibited by the League of Nations and the peninsula was regularly inspected by the Germans to insure compliance. However, extensive "coal and storage cellars" were permitted, and it was such which comprised the fortifications. The most essential part of the defenses was the contingent of men there. Unlike the Texians at the Alamo who realized they were going to die only after reinforcements from Goliad failed to arrive and the decision was made not to break through Santa Anna's encirclement, the Polish defenders of Westerplatte realized that when the German invasion began, they were

doomed. It is interesting to note the keen competition which existed to gain the supreme honour of a posting to Westerplatte. Perhaps "no bastard ever won a war by dying for his country" but the defenders of the Alamo and those of Westerplatte consciously chose their deaths as an acceptable price for wreaking a bloody vengence on the enemies of their people.

Ever since the abysmal failure of the Maginot Line in 1940, it has been taken for granted that any strategy based on even the partial use of fixed defenses is absurd. I question this view. Historically fixed defenses have proved more effective as islands rather than as flankable dikes. The Maginot Line was clearly designed as a dike, as was the Great Wall of China, and both proved failures. It is unfortunate that the dike-like tactics of trench warfare had proved so effective in World War I. Otherwise, the French would undoubtedly have noted that they were basing their 1940 defense on an historically fragile strategy. Dikes generally can withstand force only from the front, as the Persians (finally) discovered at Thermopolae. If the dikes are sufficiently narrow and thick, however, they many be effective islands and very difficult to outflank. It was conceded by the panzer innovator, von Manstein, that Germany absolutely could not have taken the Sudentenland defenses in 1938 *had they been used*. This brings up another interesting point. An effective system of fixed defenses is very much dependent on the will of the people using them.

Historical examples, modern as well as ancient, of successful use of constructed defensive positions can be given ad infinitum. Among the crusading orders, the Templars and Hospitalers early discovered that they could maintain an effective Christian presence in the Holy Land only by

concentrating a large percentage of their forces in a number of strongly fortified castles. This gave them sufficient nuisance value to cause concessions by the Muslim leaders. Most of the military disasters to the orders were the result of their frequent willingness to strip their castle defenses and join the crusader barons in massive land battles against numerically overwhelming odds--as at Hattin. For over a thousand years, some of the Christian peoples in the Near East, e.g., the Armenians and the Maronites maintained their very identity by mountain fortifications.

It is interesting to note that one of the crusader fortresses--Malta--never fell to the Muslims and was only taken (by treachery) by Napoleon in 1798. In the Second World War, the connection between the resistance of Malta and the ultimate destruction of the Afrika Korps is well remembered. Even light, hastily constructed defenses, manned by people who do not know they are supposed to surrender when surrounded, can be extremely effective in slowing down the enemy advance, as proved by the 101'st Airborne during the Battle of the Bulge.

In the examples above, there seem to be some common points. First of all, fortified defense gives a ready means of increasing the ratio of the Lanchester coefficients in favour of the Blue side. One natural advantage to this type of defense is the fact that the defender can increase his Lanchester attrition ratio by a policy of construction over a period of time. This may be a more fruitful policy than placing all one's hopes on increasing ones Lanchester ratio by the design of new weapons systems.

Secondly, fortified defense should rely on adequate stores of supplies located within the "fortress perimeter." It should be assumed by the defenders that they will be completely

surrounded by the enemy for long periods of time. (In their fortress at Magdeburg, the Teutonic Knights always kept ten year's provisions for men and horses.)

Thirdly, fortified defense is a task best undertaken by well trained professionals with strong goup loyalty.

Fourthly, fortified defense is most effective when there are allied armies poised to strike the enemy at some future time and place. The fortress and the mobile striking force complement each other in their functions. The function of the fortress is to punish, harass and divide the enemy and to maintain a presence in a particular area. In general, however, offensive activities must be left to the mobile forces. The deployment of enemy forces to take fortified positions will weaken their ability to withstand mobile offensive operations.

Let us now examine modified versions of (2.3.4) and (2.3.5)

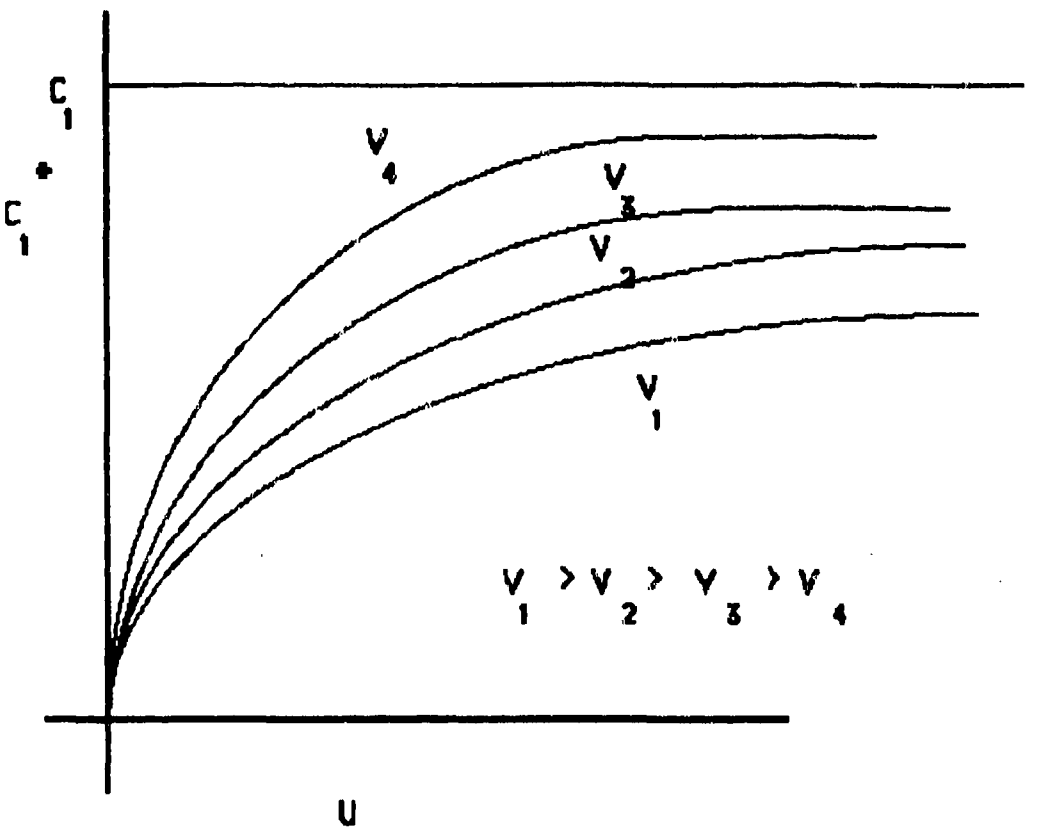
$$(2.3.20)\quad du/dt = - c_1^* v$$

and

$$(2.3.21)\quad dv/dt = - c_2^* u.$$

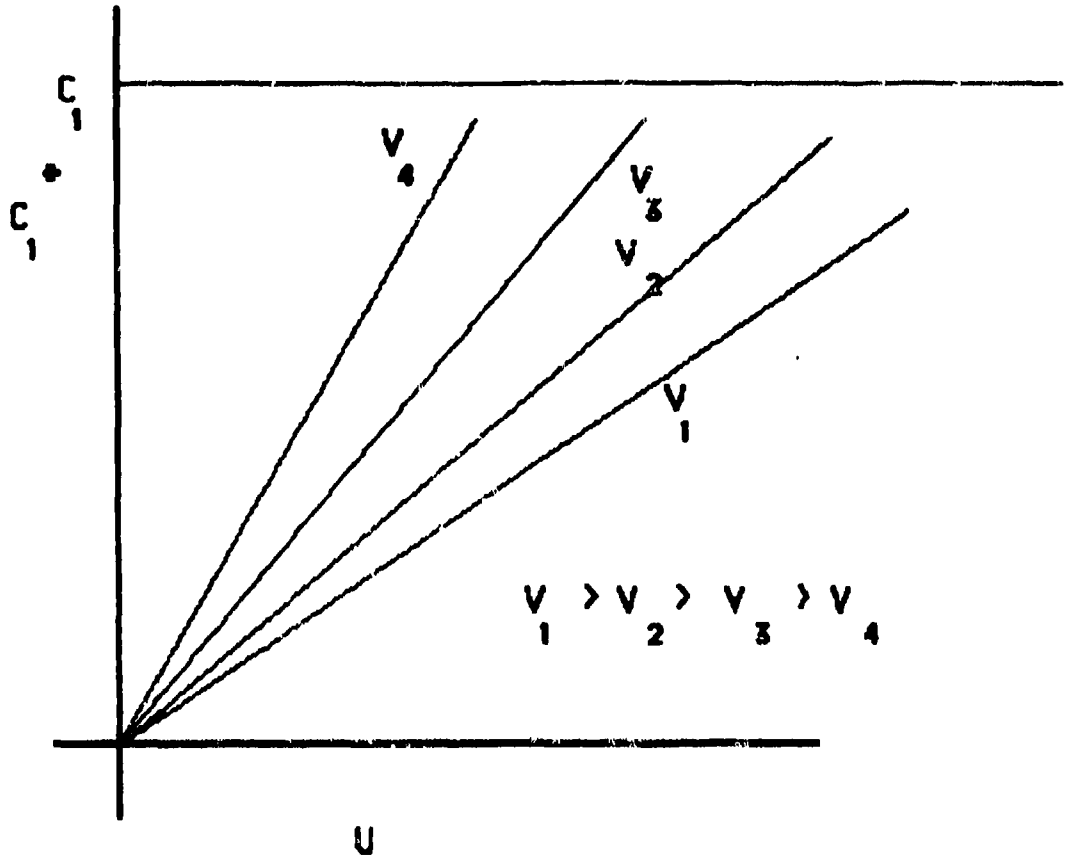
Here the attrition to Blue coefficient is taken to be variable $c_1^* = c_1^*(u,v)$ and is demonstrated graphically in Figure 5.

Figure 5



In the above, we assume that c_1^* never exceeds c_1, the attrition constant corresponding to nonfortified combat. Clearly, the functions c_1^* and c_2^* are functions of the manner in which the fortress has been constructed. It may be desirable to design the fortifications so that c_1^* is small, even at the expense of decreasing c_2^*. Generally, one might assume that c_2^* is close to the nonfortified attrition rate of u against v, since the defenders will have removed potential cover for the Red side. In fortress defense, the solution in time is likely to be important, since a primary objective is to maintain a Blue presence for as long as possible. We consider a linear approximation to the v-level curves of $c_1^*(u,v)$ in Figure 6.

Figure 6



Then we would have

(2.3.22) $du/dt = - g(v)uv - c_1^{**} u$

where $c_1^{*}(u,v) = g(v)u$ and c_1^{**} is the Blue coefficient of internal attrition. (We notice that this analysis has moved us, quite naturally to an undirected fire model for the defender's losses. The model thus derived is essentially that used earlier for the Alamo.) We might reasonably expect that the besieging forces would maintain more or less a constant number of troops in the vicinity of the redoubt. Hence we would expect

(2.3.23) $dv/dt = - c_2^{*} u - c_2^{**} v + P(u,v) = 0,$

where P(u,v) is the rate of replacement necessary to maintain

constant v strength and c_2^{**} is the Red coefficient of internal attrition. We might expect that $c_2^{**} \gg c_1^{**}$, since inadvertent self-inflicted casualties are a well known problem for the besieging force. Then

(2.3.24) $u(t) = u_0 \exp[-(g(v)v + c_1^{**})t]$.

The enemy attrition by time t is given by

(2.3.25) $\int_0^t P(u,v)dt =$

$$c_2^{**} tv + c_2^{*} u_0 (1-\exp[-(g(v)v+c_1^{**})t])/(g(v)v+c_1^{**}).$$

If the Blue defense can hold out until $u = \alpha u_0$ (where $0<\alpha<1$), then the time till the end of resistance is given by

(2.3.26) $t^* = -\ln(\alpha)/(g(v)v + c_1^{**})$.

We have, then that the total losses to the Red side by the time the defense falls is given by

(2.3.27) $[c_2^{*} u_0(1-\alpha) - c_2^{**} v \ln(\alpha)]/(g(v)v + c_1^{**})$.

It is interesting to note that if $c_2^{**}=0$, then the minimization of Red casualties appears to be consistent with the minimization of t^*. This might indicate that an optimum strategy for Red is to overwhelm the Blue fortifications by sheer weight of numbers. This would not be true if beyond some value of v, $d[g(v)v]/dv \leq 0$, implying that beyond a certain strength, additional Red forces would actually impair Red's ability to inflict casualties on the Blue side. As a matter of fact, the history of fortified defense seems to indicate that such a "beginning of negative returns" point in the v space does exist. It is generally the case for the besieging force that $c_2^{**} \gg 0$ and that it is increasing in v. This is particularly true if the besieged forces are able from time to time to conduct

carefully planned "surprises" in order to encourage increased confusion and trigger happiness on the part of the besiegers.

In the heterogeneous force model for fortified defense, we have

(2.3.28) $du_j/dt = - \Sigma_{i=1 \text{ to } n} k_{ij} g_{ij}(v_i)\, v_i u_j - c_{1j}^{**} u_j$

(2.3.29) $dv_i/dt = - \Sigma_{j=1 \text{ to } m} l_{ji} c_{2ji}^{*} u_j - c_{2i}^{**} v_i$.

The size of the j'th Blue subforce at time t is given by

(2.3.30) $u_j(t) = u_j(0) \exp[-t(\Sigma_{i=1 \text{ to } n} k_{ij} g_{ij}(v_i) v_i + c_{1j}^{**})]$

The total attrition to the ith enemy subforce at time t is given by

(2.3.31) $\int_0^t P_i(u,v)d\tau = \Sigma_{j=1 \text{ to } m} l_{ji} c_{2ji}^{*} u_j(0) \times$

$\int_0^t \exp[-\tau(\Sigma_{i=1 \text{ to } n} k_{ij} g_{ij}(v_i)\, v_i + c_{1j}^{**})]\, d\tau + c_{2i}^{**} t v_i$

$= \Sigma_{j=1 \text{ to } m} l_{ji} c_{2ji}^{*} u_j(0)\{1-\exp[-t\Sigma k_{ij} g_{ij}(v_i)\, v_i]\}/$

$\{\Sigma k_{ij} g_{ij}(v_i)\, v_i + c_{1j}^{**}\} + c_{2i}^{**} t v_i$.

Suppose that the effectiveness (at time t) of the Blue defender is measured by

(2.3.32) $T(t) = \Sigma_{j=1 \text{ to } m} a_j u_j(t)$,

where the a_j are predetermined relative effectiveness constants. If we assume that the fortress is lost when the effectiveness is reduced to some fraction α of its initial value, i.e., when

(2.3.33) $T(t) < \alpha T(0)$,

then we can use (2.3.31), in straightforward fashion, to solve for the time of capture.

The above model gives some indication of the power of the simple Lanchester "laws" in analyzing a "what if?" scenario. It

is, in large measure, the lack of "gee-whizziness" of Lanchester's models which renders them such a useful device to the applied worker. Generally speaking, after a few hours of self-instruction, a potential user can bring himself to the level of sophistication where he can flowchart his own wargame or other form of scenario analysis.

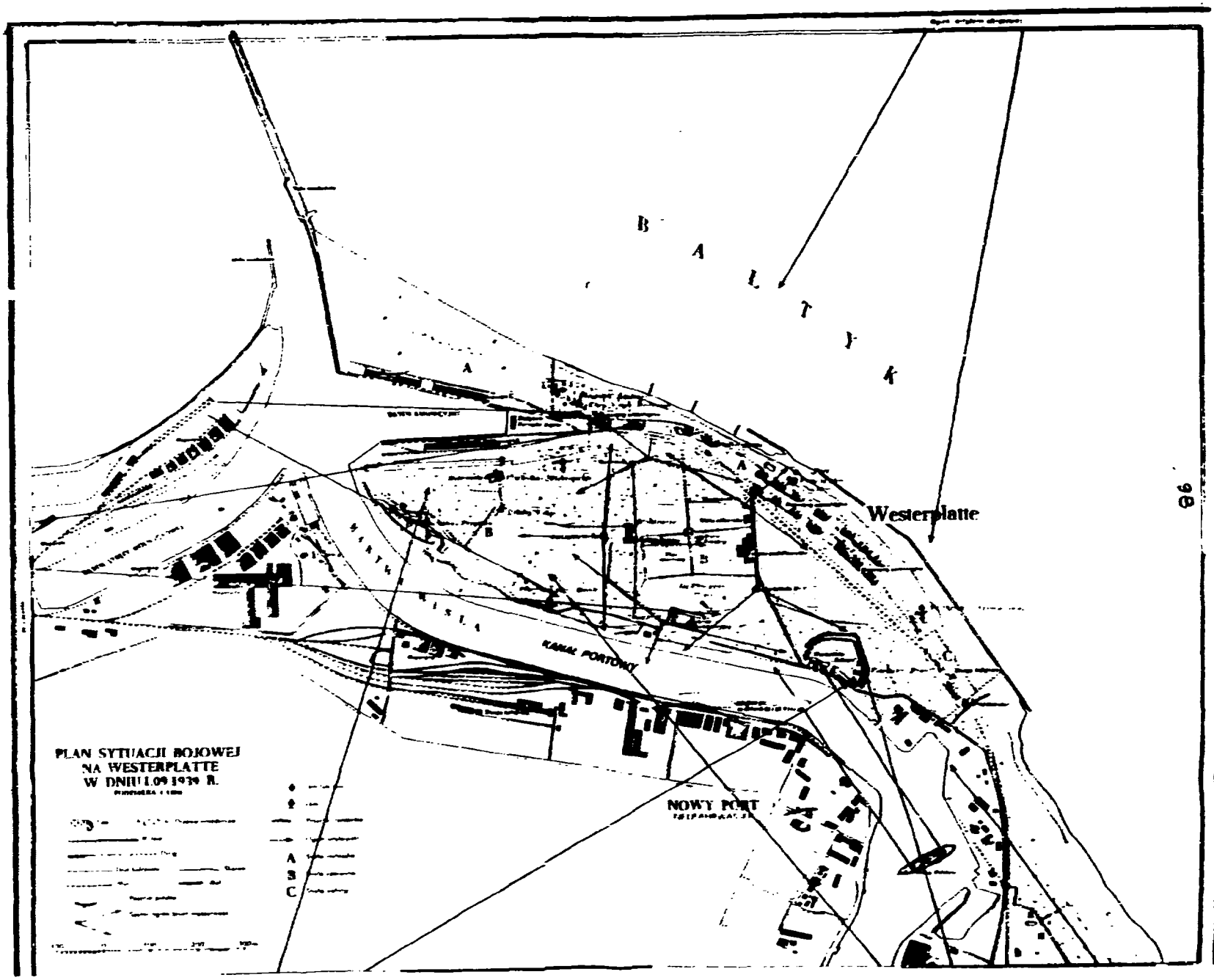
References

1. Engel, J.H. (1954)."A Verification of Lanchester's Law," *Operations Research*, v. 12,pp.344-358.

2. Lanchester, F.W. (1916) "Mathematics in Warfare," reprinted in *The World of Mathematics*, (1956),v.4, J.R. Newman, ed., pp. 2138-2157, New York: Simon and Schuster.

3. Thompson, J.R. (1979) "An Example of Data-Poor Model "Validation,"" in *Decision Information*, Tsokos,C.P. and Thrall,R.M.,eds., Academic Press,pp. 405-408.

4. Thompson, J.R. (1979) "An Argument for Fortified Defense in Western Europe", in *Decision Information*, Tsokos, C.P.and Thrall, R.M.,eds.,Academic Press, pp. 395-404.


B
A
L
T
Y
K
Westerplatte
KANAŁ PORTOWY
NOWY PORT
PLAN SYTUACJI BOJOWEJ
NA WESTERPLATTE
W DNIU 1.09.1939 R.
A
B
C

Section 4. Predation and Immune Response Systems

Let us consider Volterra's predator-prey model and some consequences for modeling the human body's anti-cancer immune response system. For the classical shark-fish model, we follow essentially Haberman [1977]. Suppose we have predators, say sharks, whose numbers are indicated by S, who prey on, say fish, whose numbers are indicated by F. In the 1920's, it was brought to the attention of Volterra that there appeared to be a periodic pattern in the abundance of certain food fish in the Adriatic, and that this pattern did not appear to be simply seasonal. Volterra attempted to come up with the simplest logical explanation of this periodicity.

We might suppose that the probability a typical shark gives birth to another shark (for reasons of simplicity we treat the sharks as though they were single cell creatures) is given by

(2.4.1) $\Pr(\text{birth in } [t,t+\Delta t]) = [aF]\,\Delta t.$

Here the assumption is that the probability of reproduction is proportional to the food supply, i.e., to the size of the fish population.

The probability a shark dies in the time interval is considered to be a constant $k\Delta t$. Thus, the expected change in the predator population during $[t,t+\Delta t]$ is given by

(2.4.2) $E[\Delta S] = S[aF-k]\Delta t.$

As we have in the past, we shall assume that for a sufficiently large predator population, we may treat the expectation as essentially deterministic. This gives us the differential equation:

(2.4.3) $dS/dt = S[\lambda F-k].$

Similarly the probability that a given fish will reproduce in

[t,t+Δt] minus the probability it will die from natural causes may be treated like

(2.4.4) Pr("birth" in [t,t+Δt]) = aΔt

We have assumed that the fish have, essentially, an unlimited food supply. The death by predation, on a per fish basis, is obviously the number of sharks multiplied by their fish eating rate, c, giving the differential equation:

(2.4.5) $dF/dt = F(a-cS)$.

Now the system of equations given by (2.4.3) and (2.4.5) has no known simple time domain solution, although numerical solution is, obviously, trivial. However, let us examine the F versus S situation by dividing (2.4.5) by (2.4.3). This gives us

(2.4.6) $dF/dS = [F/(\lambda F-k)]\,[(a-cS)/S]$.

The solution to (6) is easily seen to be

(2.4.7) $F^{-k}e^{\lambda F} = Ee^{-cS}S^{a}$, with E a constant.

Now, let us use (2.4.3) and (2.4.5) to trace the path of F versus S. We note, first of all, that $F=k/\lambda$, gives an unchanging S population; $S=a/c$ gives an unchanging F population.

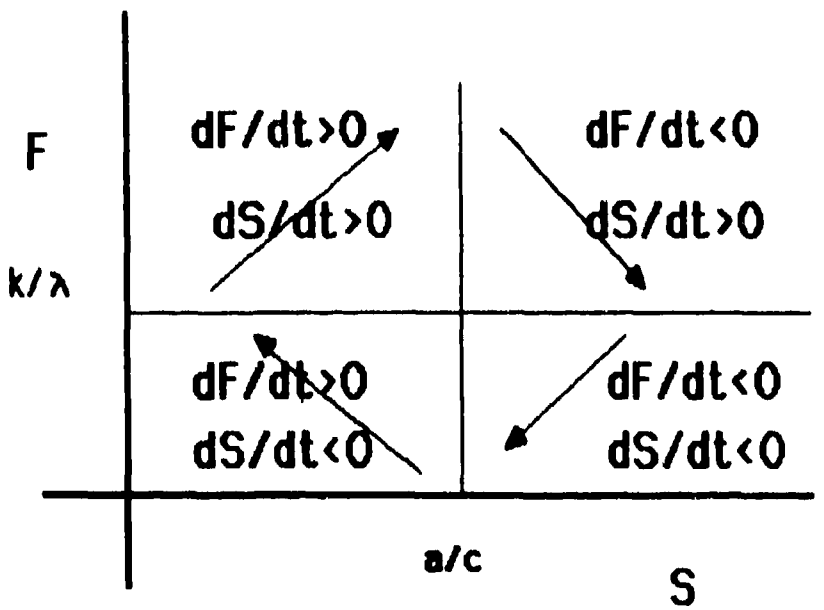


Figure 1

The consequences of Figure 1 are that the F versus S plot must either be a closed repeating curve or a spiral. We can use (2.4.7) to eliminate the possibility of a spiral. Let us examine the level curves of F and S corresponding to the common Z values in

(2.4.8) $F^{-k_e \lambda F} = Ee^{-cS_S a_{-Z}}$.

In Figure 2, we sketch the shapes of Z versus F and S, respectively, and use these values to trace the F versus S curve.

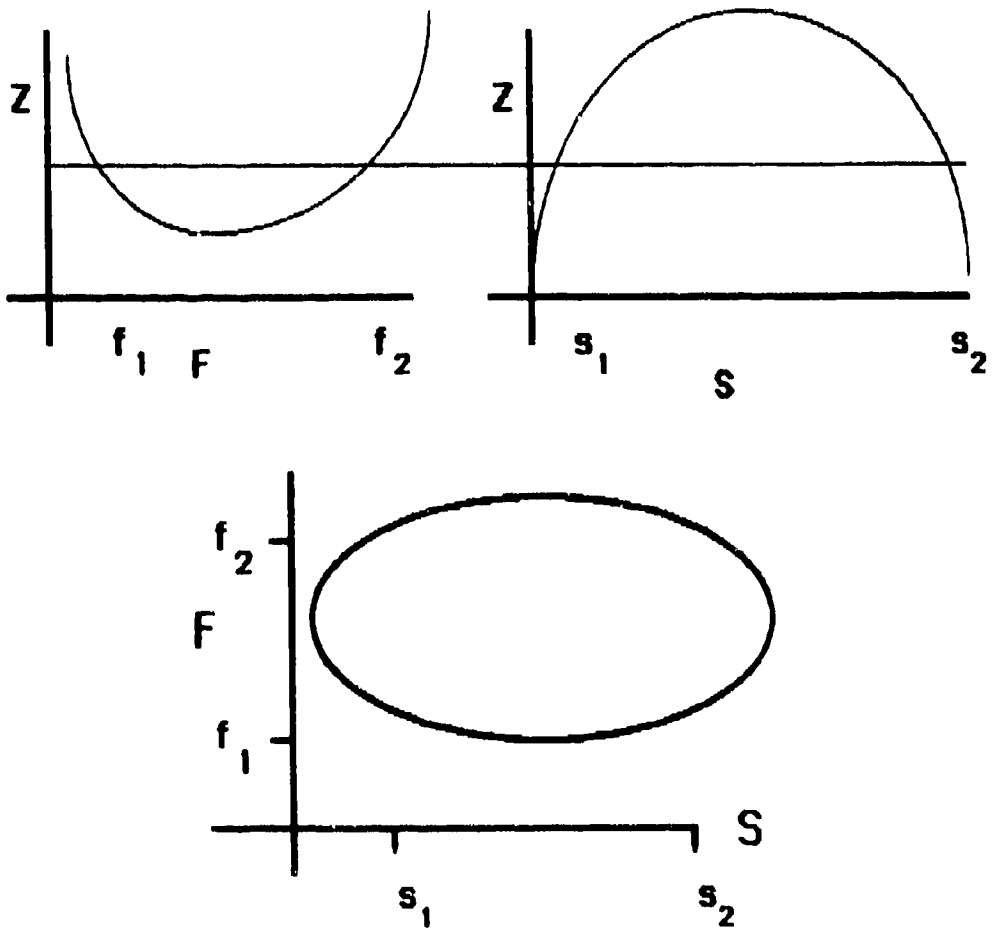


Figure 2

We note that since each value of Z corresponds to at most four

points on the F versus S curve, a spiral structure is out of the question, so we obtain the kind of closed curve which was consistent with the rough data presented to Volterra. Using Figure 1 in conjunction with Figure 2, we can sketch the time behaviour of the two populations

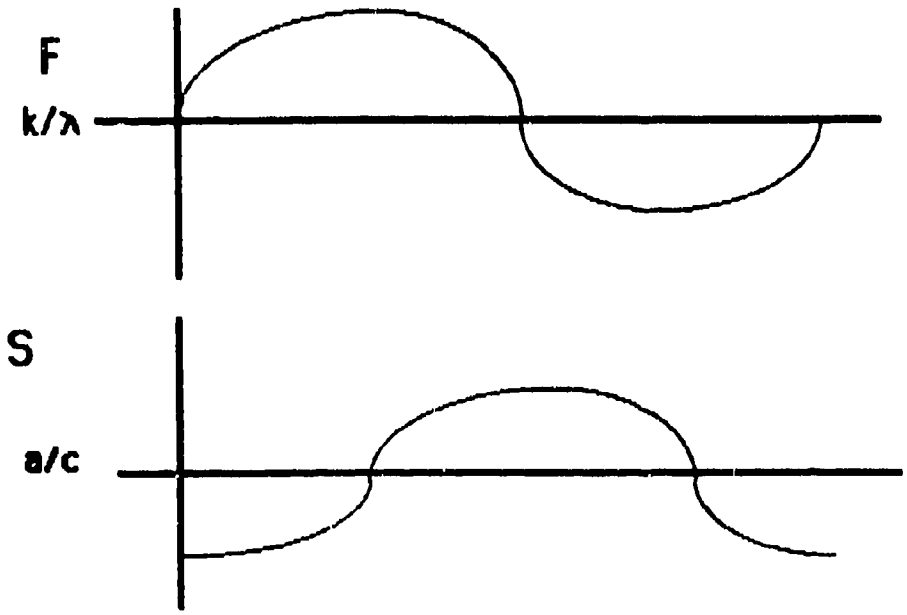


Figure 3

Here we note periodic behaviour with the fish curve leading the shark curve by "ninety degrees."

Let us now turn to an apparently quite different problem, that of modeling the body's immune response to cancer. Calling the number of cancer cells, x, let us postulate the existence of "antibodies" in the human organism which identify and attempt to destroy cancer cells. Let us call the number of these immuno entities, y, and suppose that they are given in x units, i.e., one unit of y annihilates and is annihilated by one cancer cell. Then, we can model the two populations via

(2.4.9) $dx/dt = \lambda + ax - bxy$

(2.4.10) $dy/dt = cx - bxy.$

The justification for such a model is as follows. Cancer cells are produced at a constant rate λ which is a function of environmental factors, inability of the body to make accurate copies of some of the cells when they divide, etc. a is the growth rate of the cancer cells. b is the rate at which antibodies attack and destroy the cancer cells. c is the rate of response of the antibody population to the presence of cancer cells.

Although we cannot obtain closed form solutions for the system given by (2.4.9) and (2.4.10), we can sketch a system of curves which will give us some feel as to which individuals will have immune systems which can cope with the oncogenesis process. From (2.4.10), we notice that y decreases if $dy/dt = cx - bxy < 0$; i.e., if $y > c/b$. If the inequality is reversed, then y will increase. Similarly, from (2.4.9), we note that x decreases if $dx/dt = \lambda + ax - bxy < 0$; i.e., if $y > (\lambda + ax)/b$. Let us examine the consequences of these facts by looking at Figure 4.

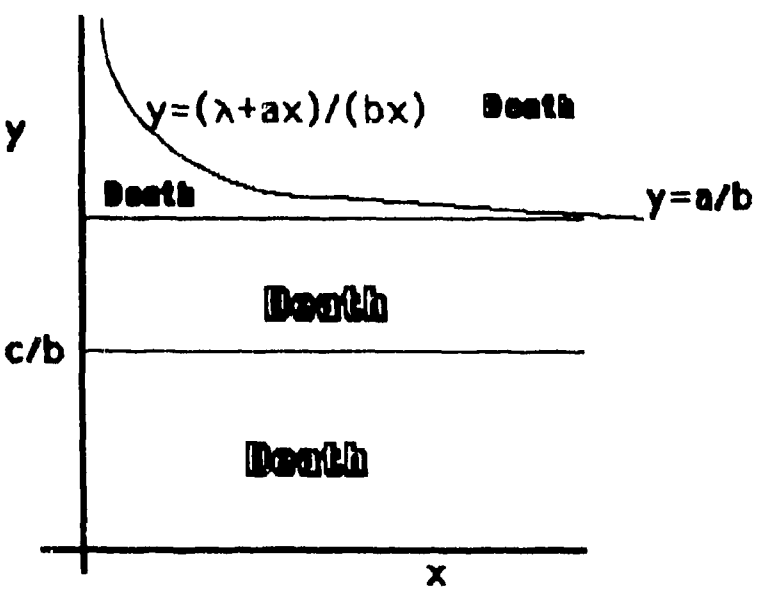


Figure 4

The prognosis here would appear to be very bad. The body is not able to fight back the cancer cells and must be overwhelmed.

On the other hand, let us examine the more hopeful scenario in Figure 5

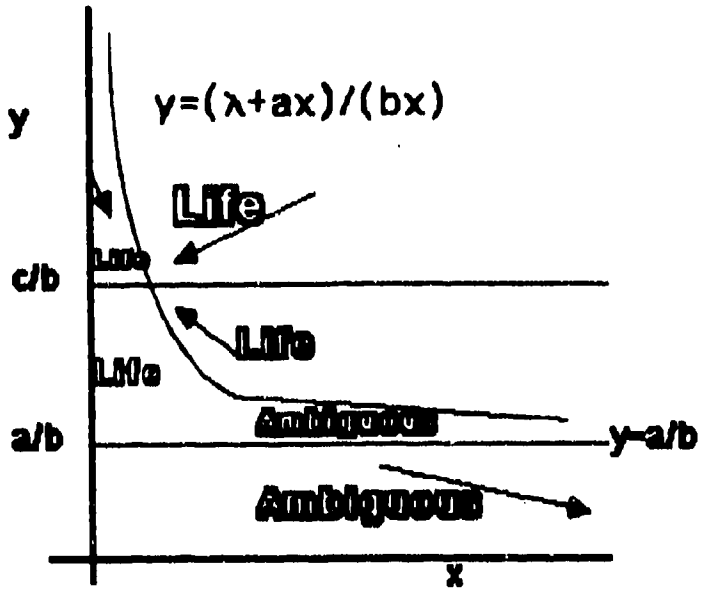


Figure 5

We note the change if c increases dramatically relative to a. We now have regions where the body will arrive at a stable equilibrium of cancer cells and antibodies. We should also note that in both Figure 4 and Figure 5, the situation of an individual who starts out with no antibodies backup at the beginning of the process is bad.

We can glean other insights from the model. For example, a large enough value of λ can overwhelm any value of c. Thus no organism can reasonably expect to have the immune response power to overcome all oncogenic shocks, no matter how big. Next, even if x is very large, provided only that we can change the biological situation to increase dramatically c, while

suppressing λ, the tumor can be defeated.

The model considered here is obviously not only hugely simplified, but purely speculative. We have, at present, no good means of measuring x and y. But it should be remembered that the model generally precedes the collection of data: generally, data is collected in the light of a model. In the case of Volterra's fish model, partial data was available because the selling of fish was measured for economic reasons. Volterra was, in short, fortunate that he could proceed from a well developed data set to an explanatory model. This was serendipitous, and unusual.

Generally speaking, we waste much if we insist on dealing only with existing data sets and refuse to conjecture on the basis of what may be only anecdotal information. If we are being sufficiently bold, then for every conjecture that subsequently becomes substantiated we should expect to be wrong a dozen times. Model building is not so much the safe and cozy codification of what we are confident about as it is a means of orderly speculation.

References

1. Haberman, Richard (1977), *Mathematical Models*, Englewood Cliffs: Prentice Hall.

Section 5. Pyramid Clubs for Fun and Profit

There are those who hold that the very formalism of the "free market" will produce good--irrespective of the production of any product or service other than the right to participate in the "enterprise" itself. One example of such an enterprise is gambling. Here, the player may understand that he is engaging in an activity in which his long run expectations for success are dim---the odds are against him. Nevertheless, he will enter the enterprise for fun, excitement and the chance that, if he only plays the game a small number of times, he will get lucky and beat the odds.

Another example of an enterprise which apparently produces no good or service is that of the pyramid club. Unlike gambling, the pyramid club gives the participant the notion that they almost certainly will "win," i.e., their gain will exceed, by a very significant margin, the cost of their participation. Let us consider a typical club structure. For the cost of $2,000, the member is allowed to recruit up to six new members. For each member he recruits, he receives a commission of $1,000. Furthermore, each of the new members is inducted with the same conditions as those of the member who inducted them. Now for each recruit made by second level members, the first level member receives a commission of $100. This member is allowed to share in these $100 commissions down through the sixth level. Generally, there is some time limit as to how long the member has to recruit his second level members--typically a year. Thus, his anticipated return is

(2.5.1) Anticipated Return = $1,000 X 6
$+[6^2+6^3+6^4+6^5]X100 = 938,400$

It is this apparent certainty of gain which attracts many to pyramid enterprises. Many state governments claim that this hope of gain is hugely unrealistic, and thus that pyramid enterprises constitute fraud. We wish to examine this claim.

Let us suppose we consider only those members of society who would become members if asked. Let us say that at any given time those who are already members will be included in the pool "y" and those who have not yet joined but would if asked are included in the pool "x". If we examine the probability that a member will effect a recruitment in time interval Δ, this appears to be given by

(2.5.2) P(recruitment in $[t,t+\Delta]$ = $k\, x/(x+y)\Delta$

where k = yearly rate of recruitment if all persons in the pool were nonmembers (e.g., k= 6).

Then we have that the expected numbers of recruits by all members in $(t,t+\Delta)$ is given by:

(2.5.3) E[number of recruits in $[t,t+\Delta]$] = $ky\ x/(x+y)\Delta$.

Now there will be an exodus from the pool given by the rate δ which is the reciprocal of the average time a member is a member (say 1 year). (You should check by an infinitessimal argument that this statement is true.)

Thus, if we replace the expectation of y by y itself, and divide by Δ, and let Δ go to 0, we have

(2.5.4) $dy/dt = kyx/(x+y) - \delta y$.

Let us make the optimistic (from the standpoint of the participants) assumption that x+y is constant. And, further, let

us consider that x and y are proportions so that y=1-x. Then we have the easily solvable (using partial fractions) equation

(2.5.5) $dy/[y(k-\delta-ky)] = dt.$

So we have

(2.5.6) $t=1/(k-\gamma)\ln[y/(k-\delta-ky)]- 1/(k-y_0)\ln[y_0/(k-\delta-ky_0)].$

Now, when dy/dt=0, there is no further increase of y. Thus, the equilibrium (and maximum) value of y is given by

(2.5.7) $y_e =(k-\delta)/k.$

For the present example, where k is 6 and δ=1, the maximum value of y is .83. y_e will only be reached at t= ∞. But it is relevant to ask how long it will take before y equals, say .82. If we assume that y_0 equals .001, a little computation shows that t(y(.82)) = 1.87 years.

Now, the rate of recruitment per member per year at any given time is given by

(2.5.8) $[dy/dt]/y = [k-\delta-ky].$

At time t=1.87, and thereafter,

(2.5.9) $[dy/dt]/y =.08.$

Unfortunately, a member who joins at t=1.87 or thereafter must replace the "6" in (1) by a number no greater than .08. Thus, the anticipated return to a member entering at this time is rather less than 938,400:

(2.5.10) Anticipated Return $\leq$1,000X.08+

$$[.08^2+.08^3+.08^4+.08^5]X100= \$80.70.$$

The difference between a pyramid structure and a bona fide franchising enterprise is clear. In franchising enterprises in

which a reasonable good or service is being distributed, there is a rational expectation of gain to members even if they sell no franchises. Potential members may buy into the enterprise purely on the basis of this expectation. Still, it is clear that a different kind of saturation effect is important. The owner of a fast food restaurant may find that he has opened in an area which already has more such establishments than the pool of potential customers. But a careful marketing analysis will be enormously helpful in avoiding this kind of snafu. The primary saturation effect is not caused by a lack of potential purchasers of fast food restaurants but by an absence of customers. On the other hand, there is little doubt that many franchising operations infuse in potential members the idea that their main profit will be realized by selling distributorships. Indeed, many such operations are *de facto* pyramid operations. Thus, it would appear to be impossible for the government to come up with a nonstiffling definition of pyramid clubs which could not be circumvented by simply providing, in addition to the recruiting license, some modest good or service (numbered "collectors' item" bronze paper weights should work nicely). The old maxim of *caveat emptor* would appear to be the best protection for the public.

The model of a pyramid club is an example of epidemic structure, although no transmission of germs is involved. Nor should the term "epidemic" be considered always to have negative connotations. It simply has to do with the ability of one population to recruit, willfully or otherwise, members of another population into its ranks.

Section 6. A Model Based Examination of AIDS: Its Causes and Likely Progression

A customary approach to the control of contagious diseases in contemporary America is via medical intervention, either by preventive vaccination or by the use of antibiotics. Historically, sociological control of epidemics has been the more customary method. This has been due, in part, to the fact that vaccines were unknown before the Nineteenth Century and antibiotics before the Twentieth Century.

In the case of some ancient peoples, a large portion of the system of laws dealt with the means of sociological control of epidemics. For example, it should be noted that the 13th, 14th and half of the 15th chapter of Leviticus (131 verses) are dedicated for the sociological control of leprosy. We might contrast this with the fact that the often mentioned dietary (*kosher*) laws receive only one chapter, the 11th, with a total of 47 verses.

The notion that epidemics can always be controlled by a shot or a pill rather than by the generally more painful sociological methods caused much human suffering even before AIDS. For example, First World medicine has largely displaced isolation as a control for leprosy in the Third World. Because the methods have been less effective in practice than hoped, we have the spectacle in some countries of three generations of a family sharing the same roof and the disease of leprosy. Only in the 1980's have we (apparently) reached the level of medical control necessary to protect individuals against the effects of leprosy. But, in some sense, we have acted for half a century as though we were in possession of an anti-leprosy technology which we

did not, in actuality, have.

In the case of AIDS, we see an even more difficult (of medical control) disease than leprosy. At present, the amount of federal funds expended on AIDS is over 15% of the total federal funding for research on all oncological diseases (of which AIDS is considered to be one). My own discussion with colleagues involved in the investigation indicates that a vaccine or a cure is extremely unlikely in the near future. Accordingly, we are confronted with a disease with a 100% fatality record and a per patient medical cost (using the present heroic intervention) in the $100,000/case range. We must ask the question of whether the present main thrust of attack can be deemed optimal or even intelligent.

Below, I will give some of the arguments used in a paper written in 1983, when the extent of the disease was much less than is the case presently. First of all, we can determine the probability that a random infective will transmit the disease to a susceptible during a time interval $[t,t+\Delta t]$.

$$\text{Prob(transmission in } (t,t+\Delta t)) =$$

$$k\Delta t \; \alpha \; \frac{X}{X+Y}$$

where
k = # contacts/time
α = prob of contact causing AIDS
X = # susceptibles
Y = # infectives

To get the expected total increase in the infective population during [t,t+Δt], we multiply the above by Y, the number of infectives.

(2.6.1) $\Delta E(Y) = Y\Pr(\text{transmission in } [t,t+\Delta t])$.

For large populations, we can assume, under fairly general conditions, that the expected total change in Y is a very nearly equal to a deterministic Y, i.e.,

(2.6.2) $\Delta E(Y) \approx \Delta Y$.

Letting Δt go to zero, this yields, immediately

(2.6.3)

$$\frac{dY}{dt} = k\alpha \frac{XY}{X+Y}$$

$$\frac{dX}{dt} = -k\alpha \frac{XY}{X+Y}$$

Now, we must allow for immigration into the susceptible population (λ), and emigration (μ) from both the susceptible and infective populations and for marginal increase in the emigration from the infective population due to AIDS (δ), from sickness and death. Thus we have the improved differential equation model

(2.6.4)

$$\frac{dY}{dt} = k\alpha \frac{XY}{X+Y} - (\delta+\mu)Y$$

$$\frac{dX}{dt} = -k\alpha \frac{XY}{X+Y} + \lambda - \mu X$$

where λ = immigration
μ = emigration
δ = aids death rate

For early stages of the disease, $X/(X+Y) \approx 1$. Accordingly, we may write the approximation:

(2.6.5) $dY/dt \approx [k\alpha - \mu - \delta]Y.$

This gives us the solution:

(2.6.6) $Y = Y(0)\exp[[k\alpha - \mu - \delta]t].$

Now, we shall use some rough guesses for some of the parameters in the equations above.

We shall assume that, absent AIDS, the total target population is 3,000,000. We shall assume that an individual stays in this population an average of 15 years (yielding $\mu = 1/(15X12) = .00556$). We will use as the average time an infective remains sexually active 10 months (yielding $\delta = .1$). To maintain the population of 3,000,000 (absent AIDS), then, we require

(2.6.7) $dX/dt = \lambda - \mu X = 0$

or $\lambda = 16,666$. Now, if we combine these figures with early death data from AIDS, we can use the approximation for Y to obtain an estimate for $k\alpha \approx .263$. Below, we show a table of predicted and

observed AIDS figures using the estimates above.

Table 1 AIDS Cases

Date	Actual	Predicted
May 82	255	189
Aug 82	475	339
Nov 82	750	580
Feb. 83	1,150	967
May 83	1,675	1,587

Now, using the somewhat smaller $k\alpha$ value of .25 and an initial infective population of 2,000, we come up with the following projections *making the assumption that things continue with the parameter values above.*

Table 2. Projections of AIDS with $k\alpha$ = .25

YEAR	CUM. DEATHS	FRACTION INFECTIVE
1	6,434	.004
2	42,210	.021
3	226,261	.107
4	903,429	.395
5	2,003,633	.738
10	3,741,841	.578
15	4,650,124	.578
20	5,562,438	.578

The fraction infective column has been given, since, in the

absence of state intervention or medical breakthrough, it is this variable which provides the (sociological) feedback for the control of the disease. Any visibility of a loathsome and fatal disease in the proportion range of one percent of the target population will almost certainly cause members of that population to consider modifying their membership in it. In the days of plague in Western Europe, one could attempt to leave centers of congested population. It would appear likely that AIDS will cause a diminution of λ and k and an increase of μ. (It is very possible that the present government health service intervention actually decreases δ and so increases the spread of the disease, but this effect is probably minor.)

Let us consider, for example, the effect of diminishing k. We note that in the early stages of the disease, an equilibrium value of $k\alpha$=.1056 is obtained. At this value, with all other parameters held constant, the total body count after 20 years is 47,848 with a fraction of infectives quickly reaching .000668. Now, let us suppose that fear reduces k to 20% of its present value, by the use of condoms and some restraint in activity. Then, the table below shows that the disease quickly retreats into epidemiological insignificance.

Table 3. Projections of AIDS with $k\alpha$ = .05

YEAR	CUM. DEATHS	FRACTION INFECTIVE
1	1,751	.00034
2	2,650	.00018
3	3,112	.00009
4	3,349	.00005
5	3,471	.00002
10	3,594	.000001

But, let us suppose that a promiscuous fraction, p, retains a $k\alpha$ value L times that of the less promiscuous population. Our model becomes:

$$dY_1/dt = k\alpha X_1(Y_1+LY_2)/(X_1+Y_1+L[Y_2+X_2])-(\delta+\mu)Y_1$$

$$(2.6.8)\quad dY_2/dt = k\alpha LX_2(Y_1+LY_2)/(X_1+Y_1+L[Y_2+X_2])-(\delta+\mu)Y_2$$

$$dX_1/dt = -k\alpha X_1(Y_1+LY_2)/(X_1+Y_1+L[Y_2+X_2])+(1-p)\lambda-\mu X_1$$

$$dX_2/dt = -k\alpha LX_2(Y_1+LY_2)/(X_1+Y_1+L[Y_2+X_2])+p\lambda-\mu X_2$$

Below, we consider the case where $k\alpha = .05$, $L = 5$, and $p = .1$.

Table 4. Projection of AIDS with $p = .10$

YEAR	CUM. DEATHS	FRACTION INFECTIVE
1	2,100	.0005
2	4,102	.0006
3	6,367	.0007
4	9,054	.0008
5	12,274	.0010
10	40,669	.0020
15	105,076	.0059
20	228,065	.0091

We notice how the presence of even a small promiscuous population can stop the demise of the epidemic. But, if this proportion becomes sufficiently small, then the disease is removed from an epidemic to an endemic situation, as we see below with p=.05 and all other parameters the same as above.

Table 5. Projections of AIDS with p = .05

YEAR	CUM. DEATHS	FRACTION INFECTIVE
1	1,917	.00043
2	3,272	.00033
3	4,344	.00027
4	5,228	.00022
5	5,971	.00019
10	8,263	.00008
15	9,247	.00003
20	9,672	.00002

The dramatic effect of a small promiscuous population may be considered in the case where 90% of the population has a kα of .02 and 10% has a kα of .32. This gives a population with an overall kα of .05. If this low value is maintained across the population, then we have seen that the disease quickly dies out. But consider the situation when the mix is given as above.

Table 6. Projections of AIDS with p = .1, kα=.02, L=16

YEAR	CUM. DEATHS	FRACTION INFECTIVE
1	2,184	.0007
2	6,536	.0020
3	20,583	.0067
4	64,157	.0197
5	170,030	.0421
10	855,839	.0229
15	1,056,571	.0122
20	1,269,362	.0182

One prediction about AIDS is that there is a "Typhoid Mary" phenomenon. That means that the actual transmission rate is much higher than had been supposed, but only a fraction of the infected develop the disease quickly. Another fraction become carriers of the disease without themselves actually developing the physical manifestations of the disease, except possibly after a long interval of time. To see the effects of such a phenomenon, let us suppose $k\alpha = .05$, but 50% of those who contract the disease have a life expectancy of 100 months instead of only 10.

Table 7. Projections of AIDS with $k\alpha = .05$ and Half of the Infectives with $\delta = .01$

YEAR	CUM. DEATHS	FRACTION INFECTIVE
1	1,064	.00066
2	1,419	.00075
3	2,801	.00089
4	3,815	.00110
5	5,023	.00130
10	16,032	.00330
15	44,340	.00860
20	115,979	.02210

Such a disastrous scenario is, naturally, made much worse as we increase the fraction of those with the long sexually active life expectancy. For example, if this proportion is 90%, we have

Table 8. Projections of AIDS with 90% Having Life Expectancy of 100 Months

YEAR	CUM. DEATHS	FRACTION INFECTIVE
1	457	.0094
2	1,020	.0013
3	1,808	.0020
4	2,943	.0028
5	4,587	.0041
10	32,911	.0260
15	194,154	.1441
20	776,146	.4754

If the Typhoid Mary phenomenon is an actuality, then the effect of AIDS is likely to be catastrophic indeed. (Note that no presence of a promiscuous subpopulation is necessary to cause this catastrophic scenario.) However, this would imply that AIDS was a new disease, contrary to the historical evidence. It seems most likely that AIDS has always been endemic in a species of Central African monkey and that its presence in the human population is of long standing. Indeed, the present entry into the United States appears to be via Haiti, which has not had significant African immigration for centuries. Since the disease has been noted in the United States, studies show the disease present in Tanzania, Uganda, Zaire, etc. These studies contain even more noise than those in the United States, which are very noisy indeed. (Also, it is interesting to note that claims have been made that the disease is frequently now of epidemic proportions in the heterosexual population. How much of this latter phenomenon is real, and how much of the real

heterosexual cases are due not to sexual contact, but other factors, e.g., zealous local medicos dispensing shots with unsterilized needles, is a matter of conjecture. If vectoring sexual vectoring into the heterosexual population is truly such an enormous problem in Africa, we need quickly to understand what the reasons are.) How likely, we must ask, is it that genetic drift in the AIDS virus would have proceeded in such widely separated populations to produce epidemics in both the United States and Central Africa at the same time? Anecdotally, a pathologist at the Texas Medical Center has informed me that some of his colleagues, nearing retirement, now recall young male patients with AIDS symptoms as long as 30 years ago, but in such occasional numbers that there was no attempt to characterize such rare occurrences in any systematic fashion.

If AIDS is not a new disease (and evidence that it is might well be investigated as an act of war by a hostile power with genetic engineering capabilities), then we ought to ask what has changed in order that an endemic disease has now reached epidemic proportions. It seems most likely that the reason is that the large contact rates (k), which characterize the frenetically homosexual communities which exist in some American cities, have never occurred before in the history of the world.

Some Suggestions From The Model

1. The most important elements in AIDS which will cause its essential elimination are:

   low value of α

   awfulness of the disease

2. As the diseased fraction of the target population increases,

k will decrease

λ will decrease

μ will increase

3. Intervention by the state in the form of a publicity campaign giving a graphically realistic assessment of the prognosis of an AIDS victim would be useful. Because of the very significant effect by a small subgroup having large numbers of potentially contagious contacts, the closing of meeting places (bath houses, etc.) where high contact rate activity takes place would be useful. If such places were closed, then the homosexual communities in a number of American cities could possibly last indefinitely. However, the resistance to such steps on the basis of civil libertarian considerations, will insure the destruction of these communities .

4. Vectoring into the heterosexual population will not be a serious problem because of the much lower level of promiscuity among straights.

5. AIDS will eliminate the target subculture, not through fatality but through fear of fatality. The ultimate "cure" of the disease will be sociological, rather than medical.

REFERENCES

1.Thompson, J.R.(1984) "Deterministic versus Stochastic Modeling in Neoplasia," in *Proceedings of the 1984 Summer Computer Simulation Conference*, pp. 822-825.

2."Update: Acquired Immunodeficiency Syndrome (AIDS) ," *United States Morbidity and Mortality Weekly Report*, Center for Disease Control, v. 32, 1983, pp. 389-391.

CHAPTER 4. SOME TECHNIQUES OF NONSTANDARD DATA ANALYSIS

Section 1. A Glimpse at Exploratory Data Analysis

Books have been written on John W. Tukey's revolutionary technique of exploratory data analysis (which is generally referred to simply as EDA), and we can only hope in a brief discussion to shed some light on the fundamentals of that subject. Moreover, the point of view that I take in this section represents my own perceptions, which may be very different from those of others. Some of the enthusiasts of EDA frequently take a philosophical position which I would characterize as being very strongly toward that of the Radical Pragmatist position in the introduction. A common phrase that one hears is that "EDA allows the data to speak to us in unfettered fashion." The "fetters" here refer to preconceived models which can get between us and the usefull information in the data. The position might be characterized by Will Rodger's famous dictum, "It isn't so much ignorance which harms us. It's the things we know that aren't so."

Whereas I believe that perceptions are always in the light of preconceived models, which we hope to modify and see evolve, there is much more to EDA than the anti-model position of some of its adherents. It is this "much more" about which I wish to speak. The digital computer is a mighty device in most quantitative work these days. Yet it has serious limitations which did not so much apply to the now discarded analog devices of the 1950's. Analog devices were very much oriented toward holistic display of the output of a model. They were not oriented toward dealing with mountains of data, nor were they particularly accurate. Digital devices, on the other hand, can be

made as accurate as we wish and handle the storage and manipulation of digitized information extremely well.

At this point in time, we have hardware which is very much more "trees" oriented than "forest" oriented. We can easily ask that this or that set of operations be performed on this or that megabyte of encoded data. But we are increasingly aware of the cognitive unfriendliness of coping with digitally processed information. Analog devices were much closer to the way the human brain reasons than are digital devices.

Perhaps what is needed is a hybridized device which combines the strong points of both the analog and digital computers. But such a hardware device will be years in bringing to a successful construction. In the mean time, what do we do? One approach might be simply to try to beat problems to death on the number cruncher. But such an approach quickly stalls. We have the computer power to obtain pointwise estimates of ten dimensional density functions using data sets of sizes in the tens of thousands. But where shall we evaluate such a density function? How shall the computer be trained to distill vast bodies of information into summaries which are useful to us? These are difficult problems and the answers will be coming in piecemeal for some time.

In the meantime, we need to cope. It is this necessity somehow to address the fact that the digital computer has outstripped our abilities to use the information it gives us that EDA addresses. Needing a good analog processor to handle the digital information and having none, a human observer is used to fulfill the analog function.

One recurring theme in science fiction has been the human who is plugged into a computer system. But the observer in EDA,

unlike the sci-fi cyborg is not hardwired into the system, is not deprived of his freewill, is in fact in control of the digital system. One present limitation of exploratory data analysis is the slow input-output performance of freewilled human observers. Thus, man-in-the-loop EDA could not be used, for example, to differentiate between incoming missles and decoys in the event of a large scale attack. EDA is exploratory not only in the sense that we can use it for analyzing data sets with which we have little experience. We should also view EDA as an alpha step toward the construction of the analog-digital hybrid computer, which will not have the slow input-output speeds of the human-digital prototype.

In the discussion below, we shall address some of the important human perception bases of EDA. Let us give a short list of some of these:

(1) The only function which can be identified by the human eye is the straight line.

(2) The eye expects adjacent pixels to be likely parts of a common whole.

(3) As points move far apart, the human processor needs training to decide when points are no longer to be considered part of the common whole. Because of the ubiquity of situations where the Central Limit Theorem, in one form or another, applies, a natural benchmark is the normal distribution.

(4) A point remains a point in any dimension.

(5) Symmetry reduces the complexity of data.

(6) Symmetry essentially demands unimodality.

Let us address the EDA means of utilizing the ability of the human eye to recognize a straight line. We might suppose that since linear relationships are not all that ubiquitous, the fact

that we can recognize straight lines is not particularly useful. Happily, one can frequently reduce monotone relationships to straight lines through transformations. Suppose, for example, that the relationship between the dependent variable y and the independent variable x is given by

(4.1.1) $y = 3e^{.2x}$

We show a graph of this relationship in Figure 1.

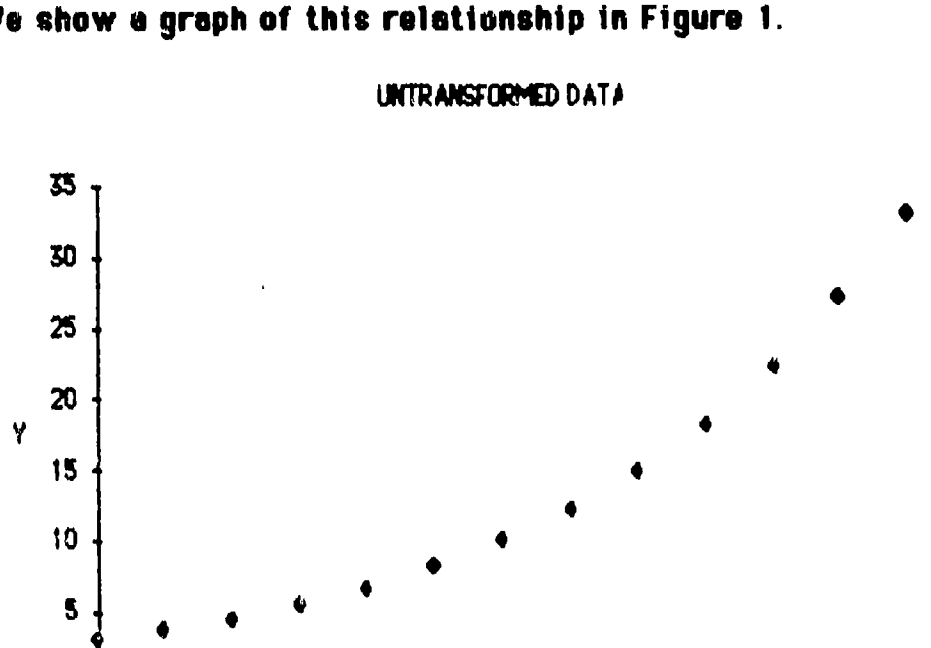


Figure 1

We can easily see that the relationship between x and y is not linear. Further, we see that y is increasing in y at a faster than linear rate. Further than this, our visual perceptions are not of great use in identifying the functional relationship.

But suppose that we decided to plot the logarithm of y against x as shown in Figure 2.

TRANSFORMED DATA

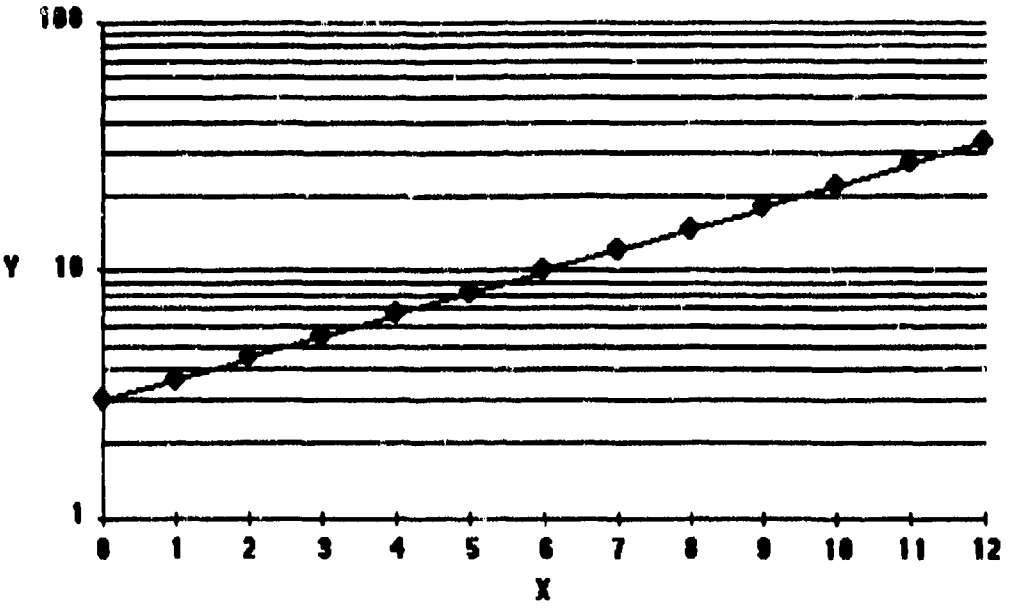


Figure 2

Now we have transformed the relationship between x and y to a linear one. By recalling how we transformed the data, we can complete our task of identifying the functional relationship between x and y. So, then, we recall that we started with an unknown functional relationship

(4.1.2) $y = f(x)$.

But then we saw that $\ln(y)$ was of the form

(4.1.3) $\ln(y) = a + bx$.

Exponentiating both sides of (4.1.3), we see that we must have a relationship of the form:

(4.1.4) $y = e^a e^{bx}$.

Once we know the functional form of the curve, we can estimate the unknown parameters by putting in two data pairs (x_1, y_1) and (x_2, y_2) and using (4.1.3) to solve:

(4.1.5) $\ln(y_1) = a + bx_1$

$\ln(y_2) = a + bx_2$.

This immediately gives the true relationship in (4.1.1).

Clearly, we will not always be so fortunate to get our transformation to linearity after trying simply a semilog plot. We might, for example, have the relationship

(4.1.6) $y = 3x^{.4}$.

In such a case, simply taking the logarithm of y will not give a linear plot, for

(4.1.7) $\ln(y) = \ln(3) + .4\ln(x)$

is not linear in x. But, as we see immediately from (4.1.7), we would get a straight line if we plotted ln(y), not versus x, but versus ln(x). And, again, as soon as the transformation to linearity has been achieved, we can immediately infer the functional relationship between x and y and compute the parameters from the linear relationship between ln(y) and ln(x).

Now it is clear from the above that simply using semilog and log-log plots will enable us to infer functional relationships of the forms

(4.1.8) $y=ae^{bx}$

and

(4.1.9) $y=ax^{b}$, repectively.

This technique of transforming to essential linearity has been used in chemical engineering for a century in the empirical modeling of complex systems in mechanics and thermodynamics. Indeed, the very existence of log-log and semilog graph paper is motivated by applications in these fields. In the classical applications, x and y would typically be complicated dimensionless "factors," i.e., products and quotients of parameters and variables (the products and quotients having been empirically arrived at by "dimensional analysis") which one would plot from experimental data using various kinds of graph

paper until linear or nearly linear relationships had been observed. But the transformational ladder of Tukey goes far beyond this early methodology by ordering the transformations one might be expected to use and approaching the problem of transformation to linearity in methodolical fashion. For example, let us consider the shapes of curves in Figure 3:

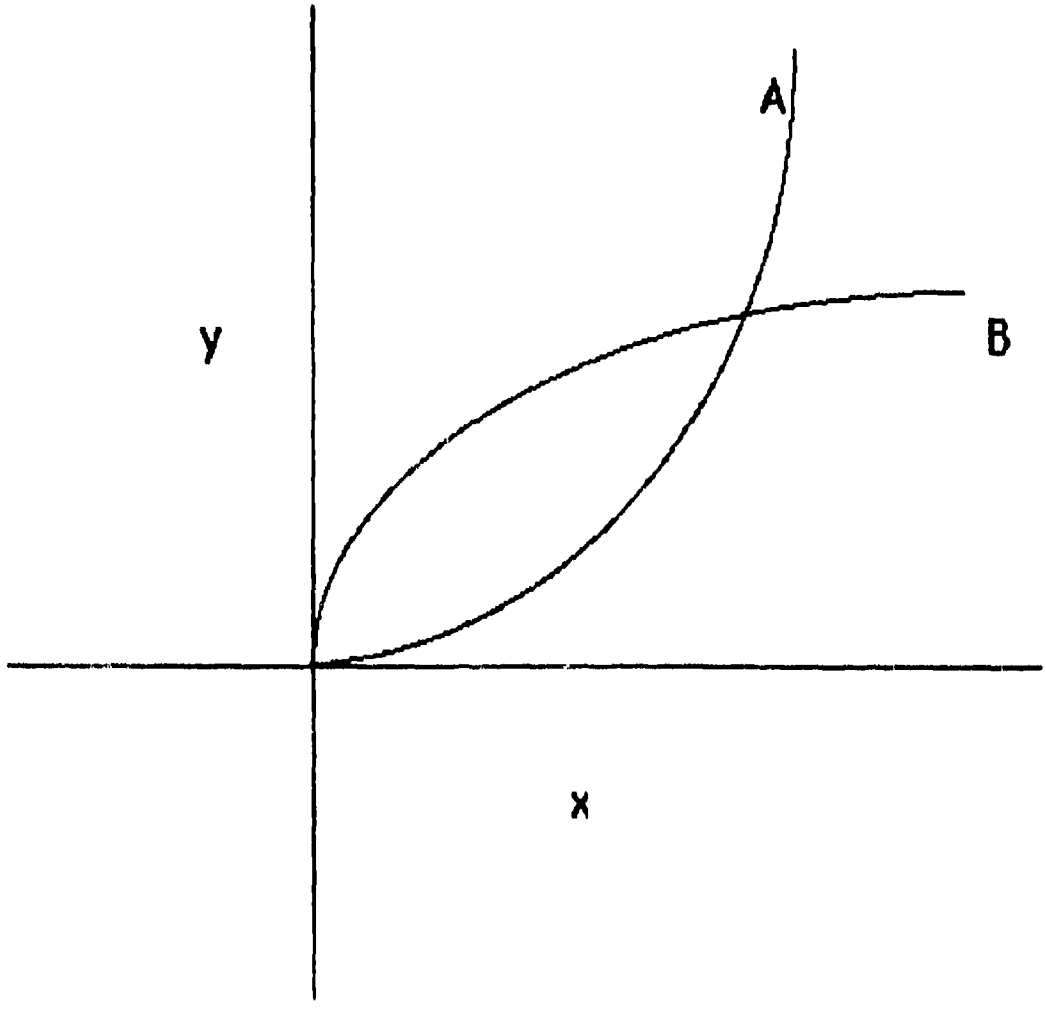


Figure 3

Now it is clear that curve A is growing faster than linearly. Accordingly, if we wish to investigate transformations which will bring its rate of growth to that of a straight line, we need

to use transformations which will reduce its rate of growth. Some likely candidates in increasing order of severity of reduction are:

(4.1.10) $y^{1/2}$, $y^{1/4}$, $\ln(y)$, $\ln(\ln(y))$.

Similarly, if curve B is to be transformed to linearity, we might try, in decreasing order of severity:

(4.1.11) $\exp(e^y)$, $\exp(y)$, y^4, y^2.

Putting the two groups of transformations together, we can build a transformational ladder:

(4.1.12) $\exp(e^y)$, $\exp(y)$, y^4, y^2, y, $y^{1/2}$, $y^{1/4}$, $\ln(y)$, $\ln(\ln(y))$.

The shape of the original y curve points us up or down the transformational ladder.

Using the transformational ladder to find more complicated functional relationships between y and x becomes much more difficult. For example, it would require a fair amount of trial

and error to infer a relationship such as

(4.1.13) $y = 4 + 2x^2 + x^3$.

Furthermore, we must face the fact that in practice our data will be contaminated by noise. Thus, uniqueness of a solution will likely evade us.

For a great many situations, the use of Tukey's transformational ladder will bring us quickly to an quick understanding of what is going on. The technique avoids the use of a criterion function and uses the visual perceptions of an observer to decide the driving mechanism.

For more complicated problems, we can still be guided by the philosophy of the technique to use the computer to handle situations like that in (4.1.13) even when there is a good degree of noise contamination. We might decide, for example to use least squares to go through a complex hierarchy of possible models, fitting the parameters as we went. So, then, we might employ

(4.1.14) $S(\text{Model(in x)}) = \Sigma(y - \text{Model})^2$.

If we have an appropriately chosen hierarchy of models, we might have the computer output those which seemed most promising for further investigation. The problem of choosing the hierarchy is a nontrivial problem in artificial intelligence. We must remember, for example, that if models in the hierarchy are overparameterized, we may come up with rather bizarre and artificial suggestions. For example, if we have 20 data points, a 19th degree polynomial will give us a zero value for the sum in (4.1.14).

Let us now turn to the second of the perception based notions of EDA: namely the fact that the eye expects continuity, that adjacent points should be similar. This notion has been used

with good effect, for example, in "cleaning up" NASA photographs. For example, let us suppose we have a noisy monochromatic two dimensional photograph with light intensities measured on a Cartesian grid as shown in Figure 4.

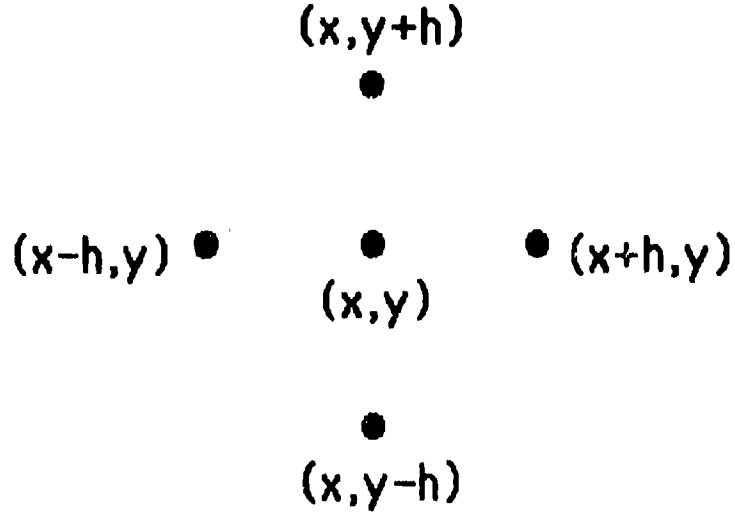


Figure 4

We might decide to smooth the intensities, via the hanning formula

$$(4.1.15)\quad I(x,y) \leftarrow [4I(x,y)+I(x-h,y)+I(x+h,y)+I(x,y-h)+I(x,y+h)]/8$$

where I(x,y) is the light intensity at grid point (x,y).

Valuable though such a smoothing device has proven itself to be (note that this kind of device was used by Tukey and his associates 40 years ago in time series applications), there is the problem that outliers (wild points) can contaminate large portions of a data set if the digital filter is applied repeatedly. For example, suppose we consider a one dimensional data set, which we will smooth using the hanning rule

$$(4.1.15)\quad I(x) \leftarrow [2I(x)+I(x-h)+I(x+h)]/4$$

At the ends of the data set, we will simply use the average of the endpoint with that of the second point. We show below the data set followed by succesive hanning smooths:

Table 1. Repeated Hanning Smooths.

Data	H	HH	HHH
1	1	1	1
1	1	1	16.61
1	1	63.44	94.66
1	250.75	250.75	235.14
1,000	500.50	375.62	313.18
1	250.75	250.75	235.14
1	1	63.44	94.66
1	1	1	16.61
1	1	1	1

We note that the wild value of 1,000 has effectively contaminated the entire data set. To resolve this anomaly, Tukey uses a smooth based on medians of groups of three down the data set, i.e., we use the rule

(4.1.16) $I(x) \leftarrow \text{Med}\ [I(x-h), I(x), I(x+h)]$

The endpoints will simply be left unsmoothed in our discussion, although better rules are readily devised. In the data set above, the smoothing by threes approach gives us what one would presumably wish, namely a column of ones.

As a practical matter, Tukey's median filter is readily used by the computer. It is a very localized filter, so that typically if we apply it until no further changes occur (this is called the 3R smoother), we will not spread values of points throughout the data set. Note that this is not the case with the hanning filter. Repeated applications of the hanning filter will continue to change the values throughout the set until a straight line

results. Consequently, it is frequently appropriate to use the 3R filter followed by one application of the hanning filter (H). The combined use of the 3RH filter generally gets rid of the wild points (3R), and the unnatural plateaus of the 3R are smoothed by the H. Far more elaborate schemes are, of course, possible. We could, if we believed that two wild points could occur in the same block of three points, simply use a 5R filter.

Below we perform a 3RH smooth on a data set of daily unit productions on an assembly line.

Table 2. Various Smooths.

Day	Production	3	3R	3RH
1	150	150		157.5
2	165	165		168.25
3	212	193		188
4	193	201		199
5	201	201		201
6	220	201		199.5
7	195	195		190.25
8	170	170		176.25
9	161	170		167.75
10	182	161		160.25
11	149	149		142.25
12	110	110		117.5
13	95	101		101.75
14	101	95		87.75
15	60	60		64.25
16	42	42		46.5
17	15	42		46.5
18	110	60		55.5
19	60	80	60	60
20	80	60		57.5
21	50	50		50
22	40	40		45

A graph quickly shows how the 3RH smooth approximates closely what we would do if we smoothed the raw data by eye.

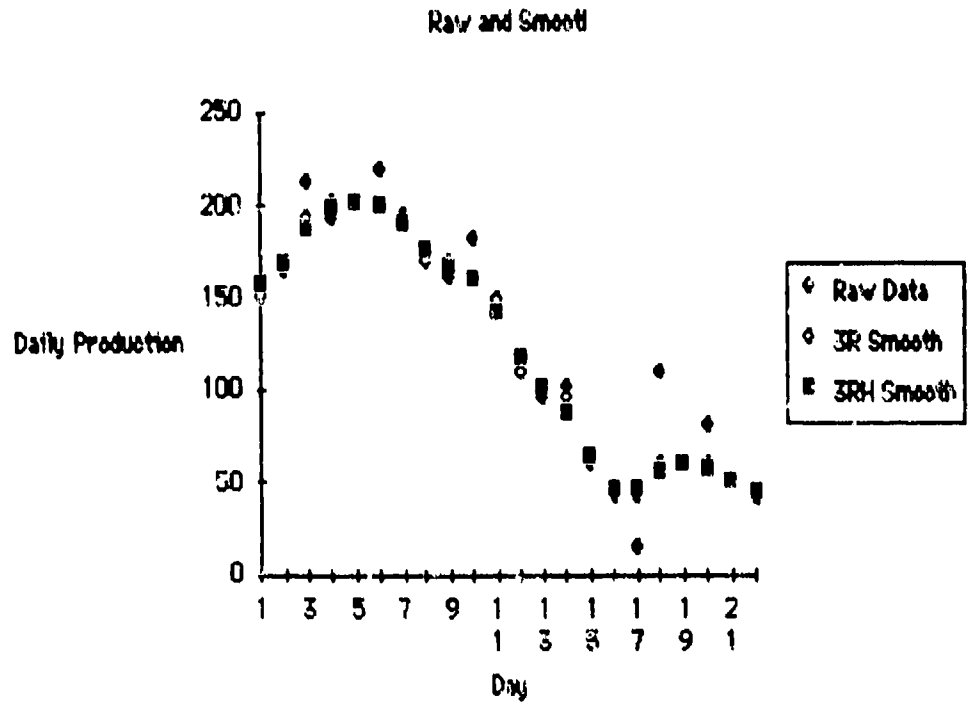


Figure 5

At this point, we should mention that all the smooths of EDA are curve fits, not derived models. We clearly find the 3RH smooth a more visually appealing graph than the raw data. But the data was measured precisely; the fluctuations really were there. So, in a sense, we have distorted reality by applying the 3RH smooth. Why have we applied it nevertheless? The human visual system tends to view and store in memory such a record holistically. Whether we smoothed the data or not, our eye would attempt to carry out more or less equivalent operations to those of 3RH. The human eye expects continuity and we do not readily perceive data digitally. The smooth gives us a benchmark (the forest) around which we can attempt to place the trees. For example, we might ask what was causing the unexpectedly low production on day 17. As we mentioned earlier,

EDA tries to assist humans to carry out the analog part of the analysis process. The 3RH smooth done on the computer very nearly reproduces the processing carried out by the human eye. In a very real sense, Tukey's deceptively simple 3RH smooth is a powerful result in artificial intelligence.

Let us now address the third point, the making of the decision that a point has removed itself from a class by extreme behaviour. We note that we have already addressed this point somewhat, since we have discussed the use of the median and hanning filters.

If we seek a benchmark by which "togetherness" of a group of points can be measured, we might decide to use the ubiquitous normal distribution. We note that for this distribution,

$$(4.1.17) \qquad P[X<x] = 1/\sqrt{(2\pi)} \int_{-\infty}^{z} \exp[-t^2/2]dt$$

where $z = (x-\mu)/\sigma$, with X having mean μ and standard deviation σ, respectively. For the normal distribution, the value z=.675 in (4.1.17) gives probability .75. By symmetry, the value z = -.675 gives probability .25. Tukey calls the corresponding x values "hinges." The difference between these standardized values is 1.35. Let us call 1.5 times this H spread (interquartile range) a step. Adding a step to the standardized hinge gives a z value of 2.7. This value of 2.7 represents the standardized "upper inner fence." The probability a normal variate will be greater than the upper inner fence or less than the lower inner fence is .007≈ one percent. Adding another step to the upper inner fence gives the "upper outer fence" (in the standardized case with mean 0 and standard deviation 1, this will give

z=4.725). The probability of a normal variate not falling between the outer fences is .0000023, roughly two chances in a million. It could be argued that a value which falls outside the inner fences bears investigation to see whether it is really a member of the group. A value outside the outer fences is most likely not a member of the group. (Note that both these statements assume the data set is of modest size. If there are a million data points, all from the same normal distribution, we would expect 700 to fall outside the inner fences and 2 to fall outside the outer fences.)

Let us examine a data set of annual incomes of a set of thirty tax returns supposedly chosen at random from those filed in 1938. Suppose the reported incomes are 700, 800, 1500, 2500, 3700, 3900, 5300, 5400, 5900, 6100, 6700, 6900, 7100, 7200,7400, 7600, 7900, 8100, 8100, 8900, 9000, 9200, 9300, 9900, 10400, 11200, 13000, 14700, 15100,16900 .

We first construct a "stem-and-leaf" plot with units in hundreds of dollars. We notice that the "plot" appears to be a hybrid between a table and a graph. In recording the actual values of the data, instead of only counts, Tukey's stem-and-leaf plot gives us the visual information of a histogram, while enabling full recovery of each data point. Here is an example where we can see both the forest and the trees.

Table 3. Stem-and-Leaf---unit 100 dollars

Depth		
2	0*	78
3	1	5
4	2	5
6	3	79
	4	
9	5	349
12	6	179
17	7	12469
13	8	119
10	9	0239
6	10	4
5	11	2
	12	
4	13	0
3	14	7
2	15	1
1	16	9

From the above stem-and-leaf plot, it is clear that certain tacit assumptions have been made. For example, we compute the "depths" from both ends of the set. Thus, a kind of symmetrical benchmark has been assumed. Let us further point to symmetry by computing the median (the average of the two incomes of depth 15 from the top and that of depth 15 from the bottom), namely 7500 dollars. The two hinges can be obtained by going up to the two averages of incomes of depth 7 and 8. Thus the lower hinge is 5350 and the upper hinge is 9600. A step is given by (9600-5350)1.5 = 6375. Thus, the two inner fences are given by -1025 and 15975. The two outer fences are given by -7400 and 22350. We note immediately one income (16900) falls outside the inner fences, but none outside the outer fences.

Let us now consider the various popular summary plots used for the income information. We have already seen one, the stem-and-leaf. Although this plot looks very much like a histogram turned on its side, we note that it shows not only the forest, but also the trees, since we could completely recover our table from the plot. In the present situation, the stem-and-leaf might be sufficient data compression. Let us consider, however, some other plots.

The "five figure summary" plot below shows the mean, hinges and extreme upper and lower incomes.

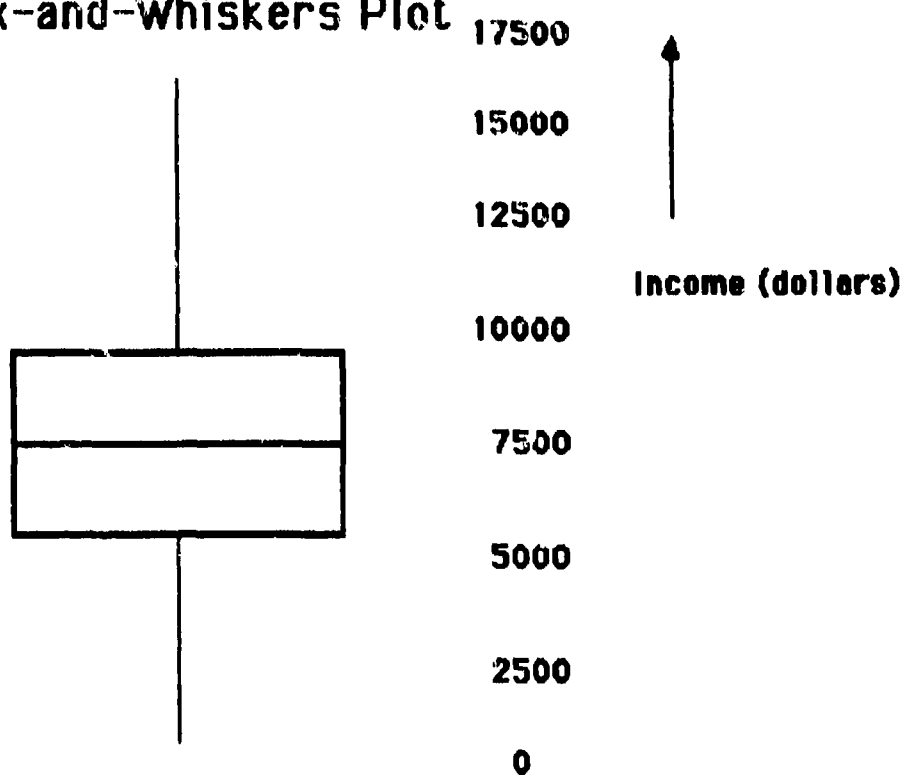
Five Figure Summary

M15h	7500	
H7h	5350	9600
1	700	16920

Figure 5

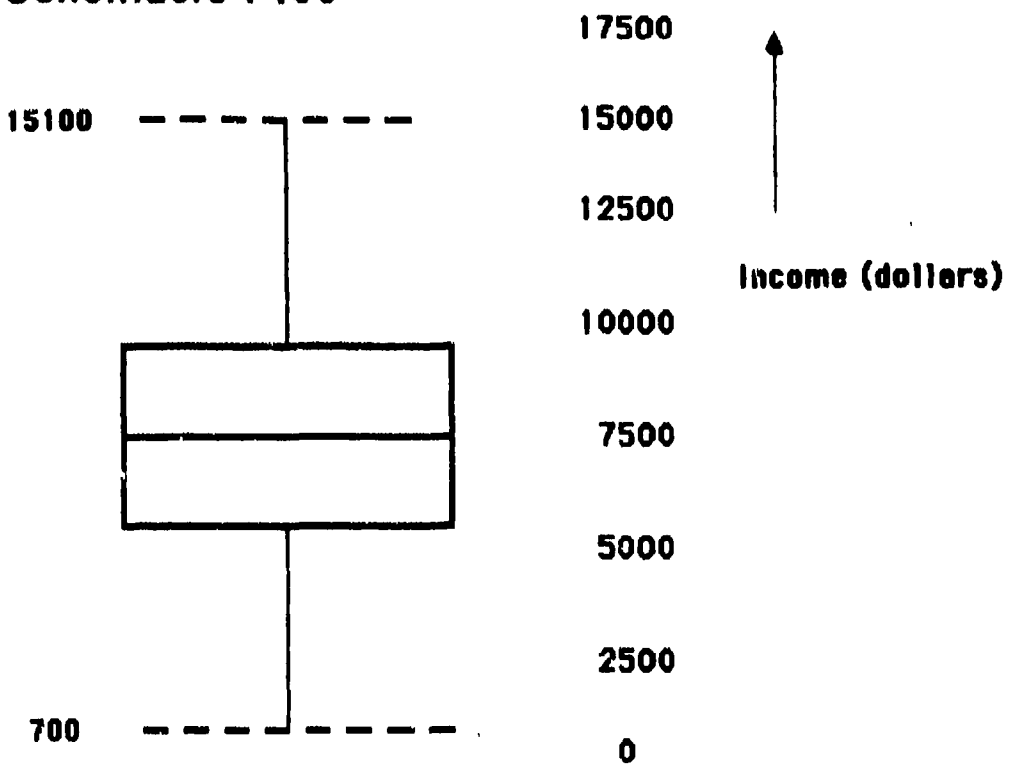
Clearly, the five figure summary is much more compressed than the stem-and-leaf. But, it draws emphasis to the supposed center of symmetry and looks at the hinges and extremal values. Naturally, as the sample becomes larger, we would expect that the median and the hinges do not change much. But the extremal values certainly will. A graphical enhancement of the five figure summary is the "box-and-whiskers" plot shown in Figure 7.

Figure 7
Box-and-Whiskers Plot



A generally more useful plot than the box-and-whiskers representation is the "schematic plot." Essentially, in this plot, the ends of the "whiskers" are the values inside the inner fences but closest to them. Such values are termed "adjacent." Essentially, then, the schematic plot replaces the extremal values with the .0035 "percentiles."

Figure 8
Schematic Plot



In the above, we seem to have a data set which is not at all inconsistent with the assumption of being all "of a piece." We might have felt very differently if, say, we had been presented with the above income data which someone had mistakenly raised to the fourth power. Going through our standard analysis, we would find values outside the upper outer fence. Yet, the data has essentially not been changed, only transformed. Before declaring points to be untypical of the group, if we believe in symmetry and unimodality, we should run through our transformational ladder until we have brought the data to a state of near symmetry. If we did this, for the example mentioned, we would arrive at something very near the original data given in Table 3, and that data set, as we have seen, does

seem to be part of the same whole.

Now it is clear that the representations of data sets discussed above are built upon the assumption of transformability to symmetry about an internal mode. If we accept this proposition, then the further use of the normal distribution as a benchmark is nontraumatic.

In the next section, we shall briefly discuss an approach, nonparametric density estimation, which does not build upon the assumption of unimodality. Obviously, such an approach must struggle with representational difficulties about which EDA need not concern itself. There is a crucial issue here. How reasonable is it to assume unimodality and symmetry, and does this assumption get better or worse as the dimensionality of the data set increases? My own view is that the problem of dealing with the pathology of outliers (extremal points which are to be discarded from membership in the data set) is not as serious as that of multimodality, and that the even more serious problem of data lying in bizarre and twisted manifolds in higher dimensional space ought to begin receiving more of our attention.

One further issue that nonparametric density estimation investigators must face is that of representation of the density function suggested by the data. For higher dimensional problems, EDA neatly sidesteps the representational issue by looking always at the original data points, rather than density contours. Let us consider two dimensional projections of a three dimensional data set generated by the routine RANDU. In Figure 9, we notice what appears to be more or less what we would expect a random set to look like. But using the interactive

routine MacSpin (D^2 Software), we can "spin" the data around the axes, to arrive at the nonrandom looking lattice structure in Figure 10.

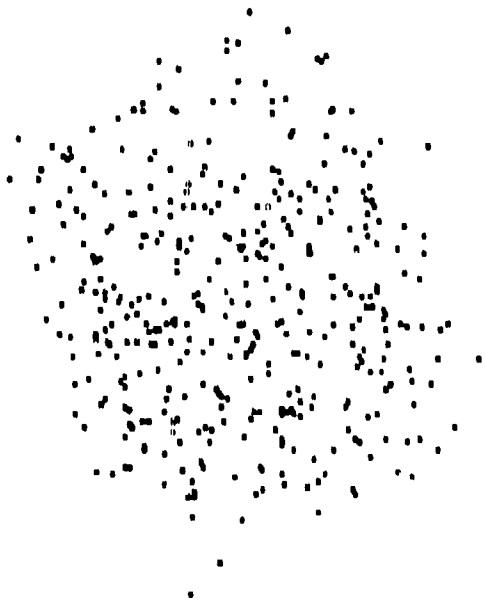


Figure 9

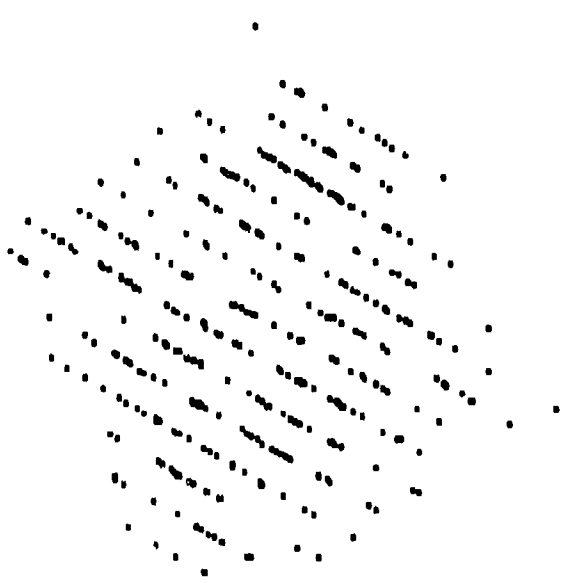


Figure 10

The human-machine interactions possible with MacSpin , a personal computer version of Tukey's PRIM-9 "cloud analysis," are truly impressive, certainly the most impressive graphics package I have yet seen for a personal computer.

Several problems of dealing always with a scattergram based analysis are obvious. For example, as the size of the data set approaches infinity, the data points will simply blacken the screen. It would appear that there are advantages to dealing with data processors that converge to some fixed, informative entity--e.g., the density function. Furthermore, whereas the automitization of such EDA concepts as the 3RH smooth are straightforward, the taking of man out of the loop with MacSpin is a very complicated problem in artificial intelligence. By opting not to use such easily automated concepts as contouring, EDA relies very much on the human eye to incorporate continuity in data analysis.

Section 2. Nonparametric Density Estimation

Perhaps the oldest procedure for looking at continuous data is that of the histogram and its precurrsor, the sample cdf. We have earlier discussed the life table of John Graunt, which gave the world its first glimpse at a cumulative distribution function. It is interesting to consider that this first approach to continuous data analysis started with an actual data set, was heuristic and preceded parametric data analysis. We see here a rather common trend in statistics, and in science more generally, namely that the search for a solution to a real problem is generally the way that important technique is developed. Although many of us spend a great deal of time trying to find applications for "useful" theory, historically the "theory in search of an application" approach is less fruitful than attempts to develop the methodology appropriate for dealing with particular kinds of real world problems.

If we know virtually nothing about the probability distribution which generated a data set, there are a number of ways we can proceed. For example, we might decide (as most do) that we will demand that the data conform to our predetermined notions of what a "typical" probability density function looks like. This frequently means that we will pull out one of a rather small number of density functions in our memory banks and use the data to estimate the parameters characterizing that density. This is an approach which has been employed with varying degrees of success for a hundred years or so.

There is a strong bias in the minds of many toward the normal (also named Gaussian or Laplacian) distribution. Thus,

we could simply estimate the mean μ and variance σ^2 in the expression

(4.2.1) $f(x|\mu,\sigma^2)=1/\sqrt{(2\pi\sigma^2)}\exp[-(x-\mu)^2/(2\sigma^2)]$.

Such a belief in a distribution as being "universal" goes back to the nineteenth century. Francis Galton coined the name "normal" to indicate this universality. He stated (1879), "I know of scarcely nothing so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'Law of Frequency of Error.' The law would have been personified by the Greeks and deified, if they had know of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason."

Galton is here discussing the practical manifestations of the Central Limit Theorem, i.e., the fact that if we sum random variables from most practical distributions, then the sum tends to a normal variate. So strong was Galton's belief in normality that in cases where the data was manifestly non-normal, he assumed that somehow it had been run through a filter before it was observed. Thus, Galton proposed such related distributions as the log-normal. Clearly the transformation to symmetry which is so important in EDA is very much in the spirit of Galton.

In most applications, it is very hard to see how the resulting data points are each in actuality, the result of a summing process which would produce normality. Nevertheless, it is a practical fact that very many data sets either are nearly normal or can be transformed to near normality by a transformation to symmetry. Galton was not naive, even less so

was Fisher. Both used the assumption of normality very extensively. Although we can get in serious trouble by assuming that a data set is normal, it seems to be a fact that we get effective normality more often than we have a right to expect.

When data is not normal, what shall we do? One approach might be to seek some sort of transformation to normality, or (in practice, almost equivalently) to symmetry. This is very much in the spirit of EDA. If the data can be readily transformed to symmetry, there is still the possibility of contamination by "outliers." These may be introduced by the blending in of observations from a second distribution, one which does not relate to the problem at hand, but which can cause serious difficulties if we use them in the estimation of the characterizing parameters of the primary distribution. Or, "outliers" may be actual observations from the primary distribution, but that distribution may have extremely long tails, e.g., the Cauchy distribution. From one point of view, EDA can be viewed as a perturbation approach of normal theory. The data is "massaged" until it makes sense to talk, for example, about a location parameter.

Nonparametric density estimation has its primary worth in dealing with situations where the data is not readily transformed to symmetry about a central mode. As such, it is much farther from normal theory than EDA. Although some (e.g., Devroye and Gyorfi) have developed techniques which are designed to handle outlier problems, the main application of nonparametric density estimation is in dealing with regions of relatively high density. Unlike both classical parametric estimation and EDA, the methodology of nonparametric density estimation is more local and less global.

For example, let us suppose that the data comes from a 50-50 mixture of two univariate normal distributions with unit variances and means at -2 and +2, respectively. The classical approach for estimating the location parameter would give us a value of roughly 0. The blind use of a trimmed mean approach would also put the location close to 0. But, in fact, it makes no particular sense to record 0 as a measure of "location." We really need to use a procedure which tells us that there is not one mode, but two. Then, using the two modal values of -2 and +2, as base camps, one can gingerly look around these local centers of high activity to get a better glimpse at the structure which generated the data.

Naturally, for low dimensional data, simply looking at scattergrams would give the user a warning that normal theory (or perturbations thereof) was not appropriate. In such cases, such EDA approaches as MacSpin are particularly useful in recognizing what the underlying structure is.

As has been noted in the section on EDA, there are problems in getting the human observer out of the loop for such procedures as MacSpin. Another problem is that in cases where there are a great number of data points, a scattergram does not converge to anything; it simply blackens the page. The scattergram does not exploit continuity in the way that nonparametric density estimation does. It makes sense to talk about consistency with a density estimator. As the data gets more and more extensive, the nonparametric density estimator converges to the underlying probability density which characterizes the mechanism which generated the data.

To get to the "nuts and bolts" of nonparametric density estimation, we recall the construction of the histogram. Let us

take the range of n univariate data points and partition it into m bins of width h. Then the histogram estimate for the density in a bin is given by

(4.2.2) $f_H(x) = (\#\text{ data points in bin containing } x)/(nh)$.

Figure 1

The graph in Figure 1 shows the kind of shape of the histogram estimator. Clearly there are disadvantages.The histogram estimator has discontinuities at the bin boundaries, and any naive attempts to use the estimator to obtain derivative information of the underlying density are inappropriate. The mean square error rate of convergence of the estimator is $n^{-2/3}$. A recent paper of Scott (1985) shows how by simply computing 16 histograms, the origin of each shifted from the preceeding h/16 to the right, and averaging point by point over each of the histograms, many of the undersirable properties of histograms are overcome, while still retaining the rapid computational speed of the histogram estimator. (For an interesting use of the histogram in bivariate systems, see Husemann (1986).)

Next to the histogram (and, significantly, the histogram is still the most used nonparametric density estimator) the most popular nonparametric density estimator is the kernel estimator, proposed first by Rosenblatt (1956) and extended and explicated by Parzen (1962). Here, the estimator at a point x is

given by

(4.2.3) $f_K(x) = \Sigma_i K((x-x_i)/h)/nh$

where K is a probability density function and the summation is over the data points $\{x_i\}$. A popular kernel here is Tukey's biweight

(4.2.4) $K(y) = (15/16)(1-y^2)^2$ for $|y| \leq 1$.

The order of convergence of the mean square error for most kernels is $n^{-4/5}$. Moreover, the procedure gives a smooth estimate as shown in Figure 2. A practical impelementation of the kernel estimation procedure (NDKER) is included in the popular IMSL library.

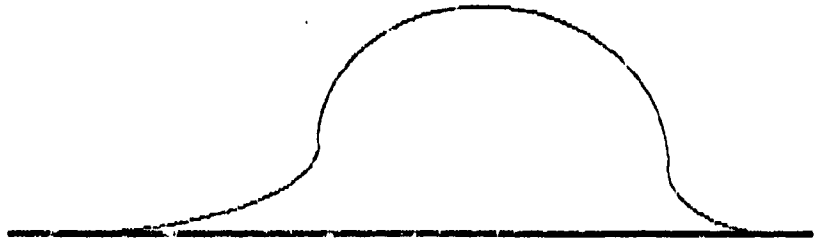


Figure 2

It is possible to use estimators of this sort to obtain derivative estimates of the underlying density. The determination of the bandwidth h can, in theory, be determined from the formula

(4.2.5) $h = n^{-1/5}[\int K^2(y)dy/(\int y^2 K(y)dy)^2]^{1/5} \times [\int (f''(y))^2 dy]^{-1/5}$.

The problem here is that we do not know f, much less f''. An approach suggested by Scott, Tapia and Thompson (1977) is to make a preliminary guess for h, use (4.2.3) to obtain an estimate for f, differentiate it, and plug into (4.2.5). The process is continued until no further change in the estimate for h is

observed. A more sophisticated approach for the selection of h has recently been given by Scott and Terrell (1986).

A more local procedure than the kernel estimator is the kth nearest neighbour kernel estimate

(4.2.6) $f(x) = \Sigma_i K((x-x_i)/d_k(x))/(nd_k(x))$,

where $d_k(x)$ is the distance from x to the kth data point nearest to it. The bandwidth parameter here is, of course, k.

Another estimation procedure is the maximum penalized likelihood approach suggested by Good and Gaskins (1971) and generalized by deMontricher, Tapia and Thompson (1975) Scott, Tapia and Thompson (1980), and Silverman (1982). In one of the simple formulations, the procedure finds the f which maximizes

(4.2.7) $J(f) = \Sigma \log f(x_i) - \alpha\int(f''(y))^2 dy$.

An implementation (NDMPLE) is given in the IMSL library. The maximum penalized likelihood approach is particularly useful in problems associated with time dependent processes (see, e.g., Bartoszyński,Brown, McBride and Thompson, 1981).

It is unfortunate that well over 95% of the papers written in the area of nonparametric density estimation deal with the univariate data case, for we now have many procedures to deal with the one dimensional situation. The problem in the higher dimensional case is very different from that with one dimensional data, as we argue below.

Suppose we are given the choice between two packets of information:

A: a random sample of size 100 from an unknown denisty

B: exact knowledge of the density on an equispaced mesh of size 100 between the 1% and 99% percentiles.

For one dimensional data, most of us, most of the time will opt

for option B. However, for four dimensional data, the mesh in option B would give us only slightly more than three mesh points per dimension. We might find that we had our 100 precise values of the density function evaluated at 100 points where the density was effectively zero. Here we see the high price we pay for an equispaced Cartesian mesh in higher dimensions. If we insist on using it, we will be spend most of our time flailing about in empty space.

On the other hand, information of packet A remains useful in four dimensional space, for it gives 100 points which will tend to come from regions where the density is relatively high. Thus they provide anchor points from which we can examine, in spherical search fashion, the fine structure of the density.

Now, we must observe that the criteria of those who deal almost exclusively with one dimensional data is to transform information of type A into information of type B. Thus, it is very wrong in nonparametric density estimation to believe that we can get from the one dimensional problem to those of higher dimensionality by a simple wave of the hand. The fact is that "even a rusty nail" works with one dimensional data. We still know very little about what works for the higher dimensional problems. Representational problems are dominant. The difficulty is not so much being able to estimate a density function at a particular point, but knowing where to look. We can, if we are not careful, spend an inordinate amount of time coming up with excellent estimates of zero. We shall discuss two of the more promising avenues of dealing with the higher dimensional problem below. The first is an attempt to extend what we have learned in density estimation in lower dimensions to higher dimensions, emphasizing graphical display. For

example we see in Figure 3 (Scott and Thompson, 1983) a display of estimated density contours using four dimensional remote sensing crop data. We note that it is quite possible to demonstrate three dimensional densities, by the use of equidensity contours. Clearly, as the value of the density function increases, we should expect to see a smaller region which satisfies the condition

(4.2.8) $f(x_1,x_2,x_3) \geq c.$

The handling of the fourth dimension, unfortunately, must be handled in a fashion asymmetrically from the other three dimensions. In Figure 3, we have employed a bar cursor at the bottom of the figure for the magnitude of the fourth variable. We note the presence of two well separated regions corresponding to a magnitude of 24 for the fourth variable. To give an idea of the scattergram alternative, we show in Figure 4, a display of the data from which Figure 3 was generated.

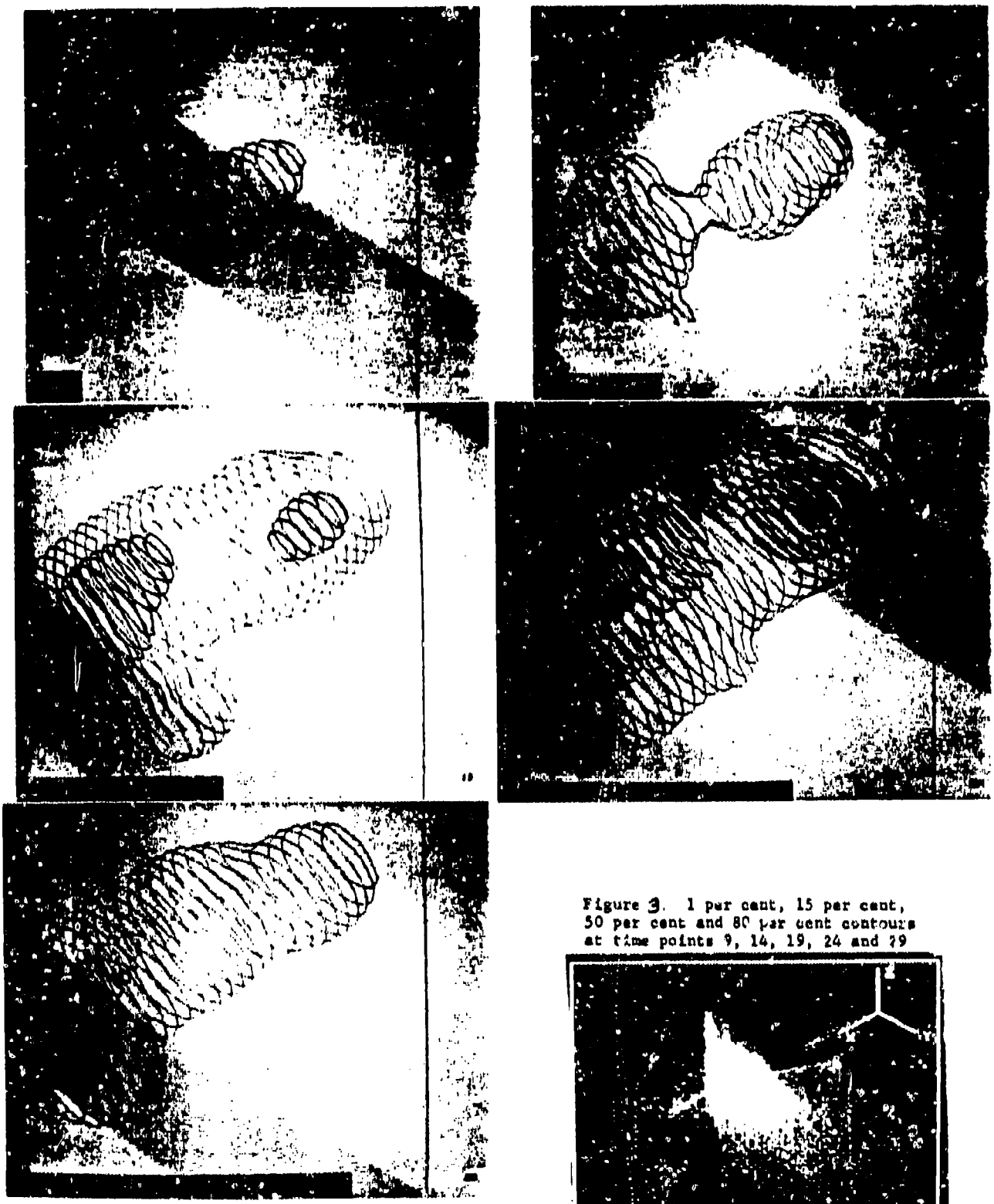
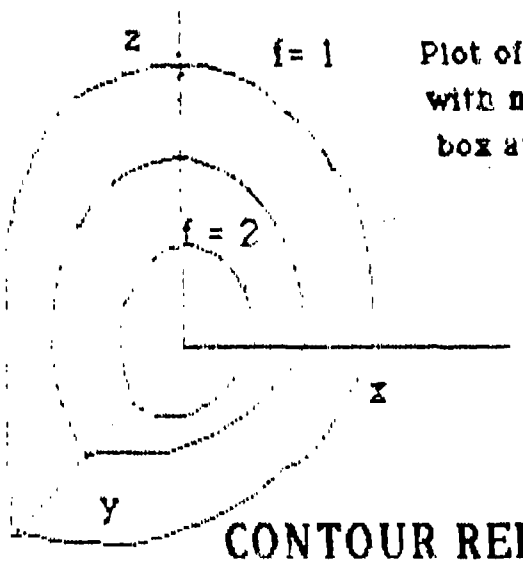


Figure 3. 1 per cent, 15 per cent, 50 per cent and 80 per cent contours at time points 9, 14, 19, 24 and 29

Figure 4

In Figure 5, we note a natural extension of the density estimation procedure above using six variables.



Plot of f contours in x,y,z with mouse set in control box at (u^*,v^*,z^*)

CONTOUR REPRESENTATION

CONTROL BOX

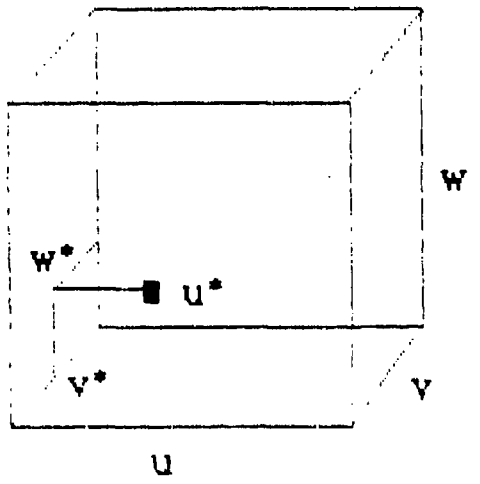


Figure 5

The contours are given in terms of three of the variables, and the magnitudes of the other three are varied using a control box.

These procedures, on the SUN 3-160 system are under investigation at Rice.

The similarities between such a density estimation approach and EDA scattergrams are clear. The problem of "fading to black" with large data sets has been eliminated. Moreover, the presence of a "man in the loop" would seem to be less than with the scattergram. The notion of a region having points of density greater than a specified amount can be automated.

A second approach (Boswell,1983, 1985) is automated from the outset. The objective of the Boswell algorithms is the discovery of foci of high density, which we can use as "base camps" for further investigation. In many situations, the determination of modal points may give us most of the information we seek. For example, if we wish to discriminate between incoming warheads and incoming decoys, it may be possible to establish "signatures" of the two genera on the basis of the centers of the high density regions.

We shall below give a brief glimpse at the simplest of the Boswell algorithms. We are seeking a point of high density, a local maximum of the density function.

(4.2.8) Algorithm 1

$x_c = x_0$

do until stopping criteria are satisfied

$x_c \leftarrow$ mean of k nearest neighbours of x_c.

In Figure 6, we sketch the result of (4.2.8) when applied to the estimation of a normal variate centered at zero with identity covariance matrix based on a sample of size 100 for dimensionality (p) through 100. If we look at the standardized

(divided by the number of dimensions) mean squared error of the estimate, we note that it diminishes dramatically as p increases to 5 and does not appear to rise thereafter.

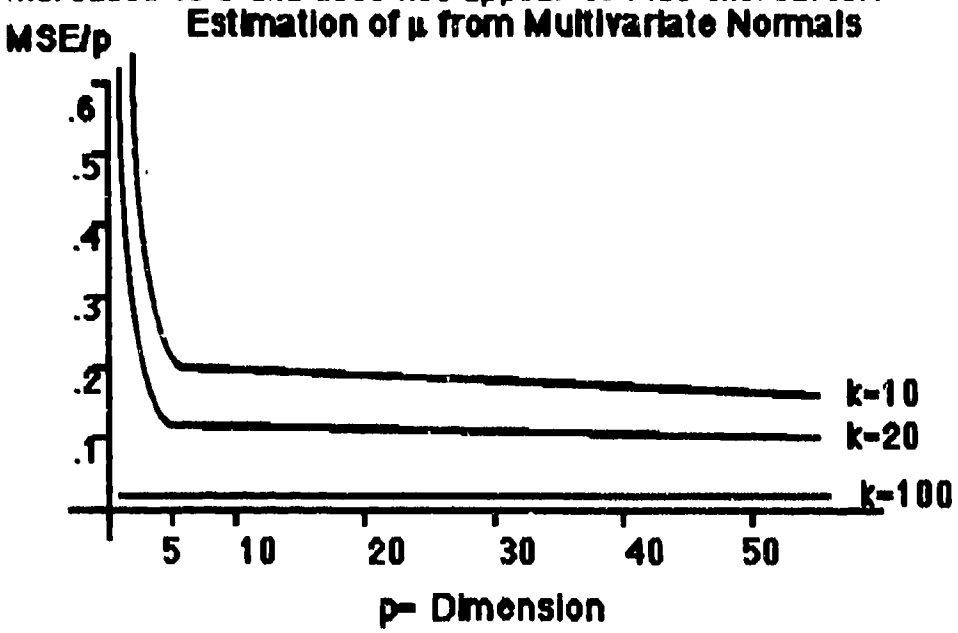


Figure 6

Naturally, we need the algorithm to deal with the more complex situation where the number of modes is large and unknown. This has been done with the Boswell approach by making multiple starts of the algorithm (4.2.8), saving the various x_c values in a file, and coalescing the estimated modes into a smaller collection.

(4.2.9) Algorithm 2

For each data point x_i set $x_c = x_i$

Perform Algorithm 1 to produce mode estimate m_i

Save m_i in a workfile

end

Analyze the set $\{m_i\}$ by cluster analytic techniques or by repeating Algorithm 2 with the $\{m_i\}$ treated as the input data set.

Algorithm 2 appears to perform reasonably well as a technique for finding the modes of mixtures of distributions (e.g., Fisher's Iris data).

In summary, the primary energies of density estimation investigators ought to be directed to the multivariate case. Nonparametric density estimation, together with EDA scattergram analysis appear to be the major contenders for handling higher dimensional data whose generating density is unknown. Many of the reasonable "nonparametric" techniques, such as rank tests, are only usable on one dimensional data. We now have the computing power available to answer some really important questions of multivariate data. For example, what price do we pay for following the usual technique of looking at low dimensional projections? Ought we to make a serious attempt to deemphasize the Cartesian coordinate system and go to spherical representations for multivariate data? When the data is not unimodal, ought we to move to multiple origin representations rather than single origin representations? How soon can we develop completely automated nonparametric density estimation algorithms for detection purposes? Can we use nonparametric density estimation as an exploratory device to get us back to algorithms based on modified normal theory?

References

(1)Bartoszyński, Robert, Brown, Barry W. and Thompson, James R. (1981), "Some Nonparametric Techniques for Estimating the Intensity Function of A Cancer Related Nonstationary Poisson Process," *Annals of Statistics*, pp. 1050-1060.

(2)Boswell, Steven B (1983), "Nonparametric Mode Estimation for Higher Dimensional Densities," Ph.D. Dissertation, Rice University.

(3)Boswell, Steven B. (1985), "Nonparametric Estimation of the Modes of High-Dimensional Densities," *Computer Science and Statistics,* L. Billard, ed., Amsterdam: North Holland.

(4)Devroye, Luc and Gyorfi, L. (1985), *Nonparametric Density Estimation: The L_1 View,* New York: Wiley.

(5) Galton, Francis (1879), *Natural Inheritance,* London: MacMillan.

(6) Good, I.J. and Gaskins (1971), "Nonparametric Roughness Penalties for Probability Densities," *Biometrika,* 255-277.

(7) Husemann, Joyce Ann (1986), "Histogram Estimators of Bivariate Densities," Ph.D. Dissertation, Rice University.

(8) deMontricher, Gilbert, Tapia, Richard A., Thompson, James R. (1975), "Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods," *Annals of Mathematical Statistics,* 1329-1348.

(9) Parzen, Emmanuel (1962), "On the Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics,* pp. 1065-1076.

(10) Rosenblatt, Murray (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics,* pp. 832-837.

(11) Scott, David W., Tapia, Richard A. and Thompson, James R. (1977), "Kernel Density Estimation Revisited," *Nonlinear Analysis,* pp. 339-372.

(12) Scott, David W., Tapia, Richard A. and Thompson, James R. (1980), "Nonparametric Probability Density Estimation By Discrete Maximum Penalized Likelihood Techniques," *Annals of Statistics,* pp. 820-832.

(13) Scott, David W. and Thompson, James R. (1983), "Probability

Density Estimation in Higher Dimensions," *Computer Science and Statistics,* J. Gentle, ed., Amsterdam: North Holland, pp. 173-179.

(14) Scott, David W. (1985), "Average Shifted Histograms," *Annals of Statistics,* pp. 1024-1040.

(15) Scott, David W. and Terrell, George R. (1986),"Biased and Unbiased Cross-Validation in Density Estimation," Department of Statistics Technical Report No. 23.

(16)Silverman, B.W. (1982), "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," *Annals of Statistics,* pp. 795-810.

(17)Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis,* New York: Chapman and Hall.

(16) Tapia, Richard A. and Thompson, James R. (1978), *Nonparametric Probability Density Estimation,* Baltimore: Johns Hopkins.

Section 3. Stein's Paradox

Suppose we wish to estimate the mean of a normal distribution with covariance matrix $\sigma^2 I$ on the basis of an observation $X=(x_1,x_2,\ldots,x_p)$. Then, if we use the loss function $L(\mu^*,\mu) = \Sigma(\mu_j{}^*-\mu_j)^2/p$, the usual estimator X has uniformly larger risk than some estimators of the form $\mu^{**}= g(X^TX)\,X$, where g is an appropriately chosen function nondecreasing between 0 and 1; i.e.,

(4.3.1) $Q(\mu^{**}) = E[L(\mu^{**},\mu)] \leq \sigma^2$

RISK VERSUS $\Sigma\mu_j$

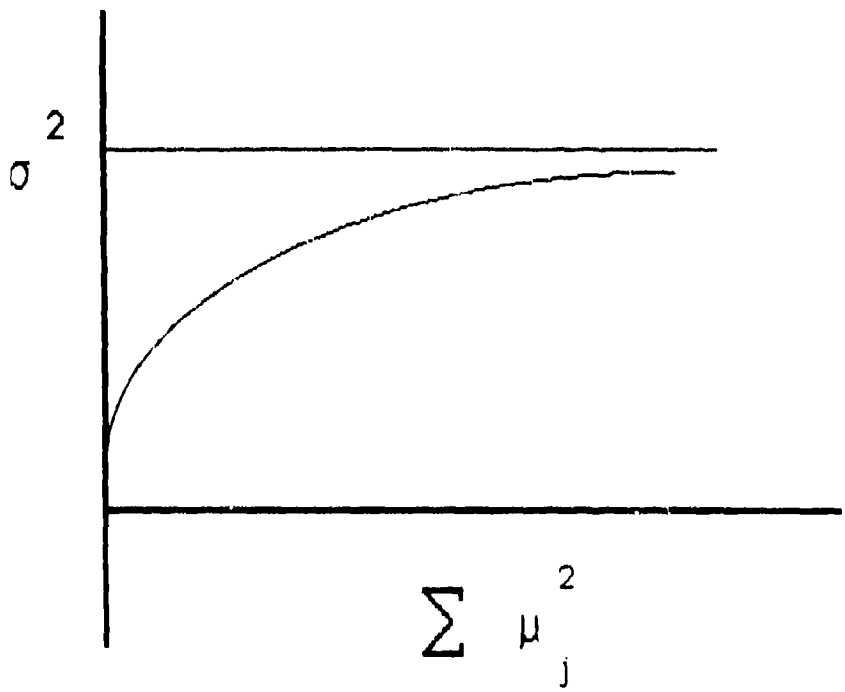


Figure 1

Lindley in commenting on a paper by Efron and Morris (1973) accordingly informs us:

....Now comes the crunch--notice it applies to the general linear model. The usual theory says x_i (maximum likelihood) is the best estimate of μ_i, but Stein showed that there is another estimate which is, for every set of μ's, better than it, when judged by the squared-error criterion except when only one or two parameters are involved. In other words, using standard criteria, the usual estimate is unsound. Further calculation (described in the paper) shows that it can be seriously unsound: with 10 parameters, quite a small number by the standard of present-day applications, the usual estimate can have five times the squared error of Stein's estimate. And remember--it can never have smaller squared error....the result of Stein undermines the most important practical technique in statistics....

The next time you do an analysis of variance or fit a regression surface (a line is aii right!) remember you are for sure, using an unsound procedure....

Worse is to follow, for much of multivariate work is based on the assumption of a normal distribution. With known dispersion matrix this can again be transformed to the standard situation and consequently, in all cases except the bivariate one, the usual estimates of the means of a multivariate normal distribution are suspect...

To get a better feel for what is happening, let us consider the one dimensional case.

Suppose we wish to estimate the mean of a random variable X on the basis of one observation of that random variable using estimators of the form

(4.3.2) $\mu_0 = aX.$

We will pick a in such a way as to minimize

(4.3.3) $Q(aX) = E[(aX-\mu)]^2 = a^2\sigma^2 + \mu^2(1-a)^2.$

Taking the derivative with respect to a and setting it equal to 0, we find the optimal a to be given simply by

(4.3.4) $a = \mu^2/(\mu^2+\sigma^2)$.

Using this a, we find that

(4.3.5) $Q(aX) = \mu^2/(\mu^2+\sigma^2)\ \sigma^2 < \sigma^2$.

This is an old result and can be found in Kendall & Stuart.

Of course, in practice, we will not have μ or σ^2 available for our finagle factor a. Still, we should ask why it is that such a factor, were it realistically available, helps us. Perhaps we get some feel if we rewrite aX as $(X/\mu)/[1+\sigma^2/\mu^2]\ \mu$. This gives us the truth---μ---degraded by a multiplier which, if μ be small (relative to σ^2), would discount, automatically, large values of X as outliers. If μ is large, (relative to σ^2), then we are left essentially with the usual estimator X. Thus, there is no paradox in the improvement of aX over X as an estimator for the one dimensional case, if we know μ and σ^2. Note, moreover, that the argument to find a did not depend on any assumption of normality, only on the existence of a finite variance.

Again, in the one dimensional case, we should address ourselves to dealing with the situation where we do not have μ or σ^2 available for our finagle factor. (I shall assume we do have σ^2 for reasons of convenience, but the argument holds if we do not have σ^2 .) In such a case, we will have available the estimator $X^2/(X^2+\sigma^2)\ X$.

But here, we generally lose our "free lunch". If the data is normally distributed, then our risk curve looks like:

Risk of $X^2/(X^2 + \sigma^2)$ X

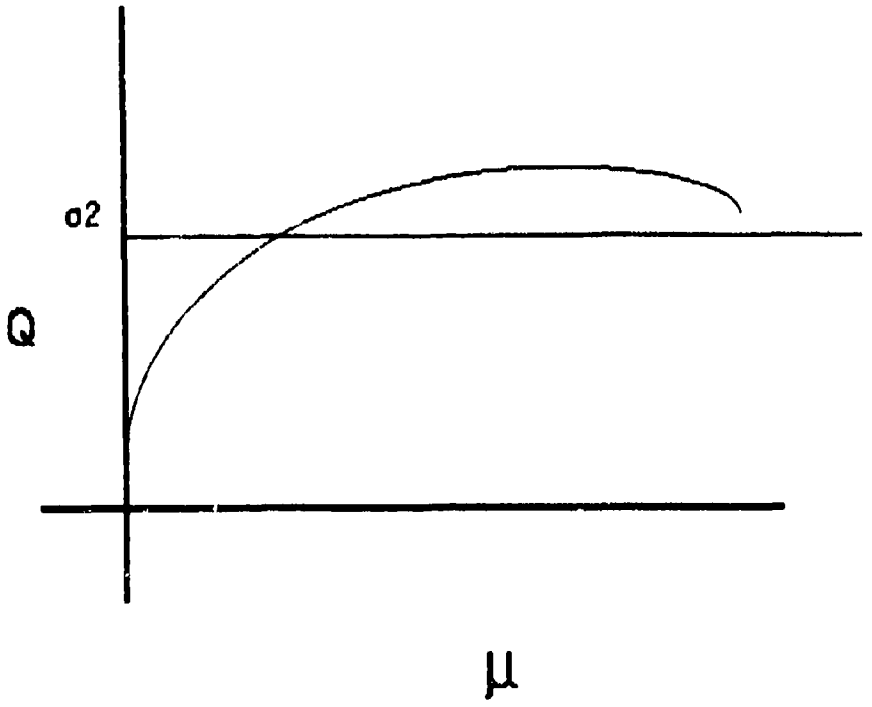


Figure 2

Let us suppose that we were allowed to use the following strategy: we will have one observation X_1 to use for the estimation of μ. But, in addition, we will have p-1 additional observations of X: $X_2, X_3, \ldots X_p$ to be used in an a of the form:

(4.3.6) $a = (\Sigma X_j^2)/(\Sigma X_j^2 + p\sigma^2)$.

Now,

(4.3.7) $\Sigma X_j^2/p \dashrightarrow \mu^2 + \sigma^2$,almost surely in p.

Thus, our finagle factor a approaches

(4.3.8) $(\mu^2 + \sigma^2)/(\mu^2 + 2\sigma^2)$ for p "large."

This would give, for large p,

(4.3.9) $Q(aX) \approx [\mu^4 + 3\sigma^2\mu^2 + \sigma^4] / [\mu^4 + 4\sigma^2\mu^2 + 4\sigma^4]\sigma^2 < \sigma^2$.

We might suspect that something in the Stein formulation may allow such a phenomenon to occur. Indeed, this is the case.

For the loss function considered

(4.3.10) $L(\mu^*,\mu) = \Sigma(\mu_j^*-\mu_j)^2/p,$

and estimators of the form:

(4.3.11) $\mu^* = X^TX / [X^TX + c\sigma^2] X,$

the risk is not dependent on the allocation of the (μ_j) for any fixed $\Sigma\mu_j^2$. (Alam and Thompson proved (1968) that, in the normal case, this estimator beats X for p>2 if $0<c<2(p-2)$.) Accordingly, we need only consider the case where $\mu = (\mu,\mu,\ldots,\mu)$. But this reduces immediately to the kind of one dimensional estimator we showed had asymptotically (in p) smaller risk than X_1.(Apparently, for the normal case, the asymptotic result starts impacting for p=3.) Thus, it is the assumption of a loss function of a *particular form*, which gives the apparent Stein improvement.

Note that for unequal weights and unknown variance, the Stein result holds, *if we know the weights in the loss function*

(4.3.12) $L(\mu^*,\mu) = \Sigma w_j(\mu_j^*-\mu_j)^2/p,$

But is it not reasonable to assume that we will frequently know the weights precisely? After all, cost functions are frequently common. So, for example, we might need to estimate $\Sigma w_j\mu_j$, where the weights are known. Note that this is the one dimensional estimation problem where we know, in the normal case, we cannot uniformly beat Σw_jX_j.

The cases where we know the weights in the Loss function

(4.3.13) $L(\mu^*,\mu) = \Sigma w_j(\mu_j^*-\mu_j)^2/p,$

are rare. Any strategy which assumes we do have precise knowledge of the weights is likely to be dangerous. Let us look at the more realistic situation where we do not have precise knowledge of the weights. Thus, let us consider the loss function

(4.3.14) $L(\mu^*,\mu,t) = \Sigma\, w_j^t\, (\mu_j^*-\mu_j)^2/p,$

where, for all $t\in T$, $\Sigma\, w_j^t = 1$, $w_j^t \geq 0$.

Let the risk $Q(\mu^*;\mu,t) = E[\, L(\mu^*,\mu,t)]$. Let the class of estimators Δ to be considered be those of the form

(4.3.15) $\mu^* = X\, f(X'X)$, where f is positive, real valued and ≤ 1.

Def An estimator μ^* is said to be *w-admissible* if there does not exist in Δ an estimator μ^{**} such that $Q(\mu^{**}) \leq Q(\mu^*)$ for all (μ,t) and for at least one (μ,t), $Q(\mu^{**})$ is strictly less than $Q(\mu^*)$.

Def An estimator is *w-minimax* if it minimizes $\sup_{(\mu,t)} Q(\mu^*;\mu,t)$ for all members of Δ.

Note that the usual estimator $(X_1, X_2, \ldots X_p)$ is w-admissible (consider the special case where $w_1^t = 1$). Moreover, $(X_1, X_2, \ldots X_p)$ minimizes $\mathrm{Max}_{\mu,t}\, Q$, i.e., is *w-minimax*. The Stein estimators cannot be w-minimax for squared loss function, since for $w_1^t = 1$, they are randomized estimates of μ_1.

In conclusion, there is no "paradox" about Stein estimation. The free lunch is due to an apparent but artificial transferral of information between the dimensions as a result of an unrealistic assumption about the loss function. Shrinkage toward an arbitrary point (without prior information), on the basis of a factor which is built up using information from

variables which are totally unrelated, which strikes most people, at first glance, as inappropriate, is indeed inappropriate.

When estimating, simultaneously, the density of mosquitoes in Houston, the average equatorial temperature of Mars, and the gross national product of ancient Persia, we ought not believe that some mathematical quirk demands that we multiply our usual (separable) estimates by a finagle factor which artificially combines all three estimates.

The above study has been given as an example of the difficulties which attend us when we attempt to make the world conform to an idealized mathematical construction, instead of the other way round. When the use of a particular criterion function yields results which are completely contrary to our intuitions, we should question the criterion function before disregarding our intuitions. At the end of the day, we may find that our intuitions were, indeed, wrong. The world is not flat, naive perceptions notwithstanding. However, the flatness of the earth was not disproved by construction of an artificial mathematical model, but rather by the construction of a model which explained real things with which the assumption of a flat earth could not cope.

References

1. Alam, K. and Thompson, J. (1968), "Estimation of the Mean of a Multivariate Normal Distribution," Indiana U. Technical Report.

2. Efron, B. and Morris,C. (1973) "Combining Possibly Related Estimation Problems," *JRSS,B,* ,v.35, pp. 479-421.

3. James, W. and Stein, C. (1961) "Estimation with Quadratic

Loss Function," *Proceedings of the Fourth Berkeley Symposium*, Berkeley: University of California Press, pp.361-370.

4. Judge,G. and Bock, M. (1978), *The Statistical Implications of Pre-Test and Stein Rule Estimators in Econometrics*, New York: North Holland.

5. Kendall, M. and Stuart, D. (1946), *The Advanced Theory of Statistics*, v. 2., New York: Hafner.

6. Thompson, J. (1968), "Some Shrinkage Techniques for Estimating the Mean," *JASA*, v. 63, pp. 113-122.

7. Thompson, J. (1968), "Accurcacy Borrowing in the Estimation of the Mean by Shrinkage to an Interval," *JASA*, v. 63, pp. 953-963.

8. Thompson, J. (1969), "On the Inadmissibility of X as the Estimate of the Mean of a p-Dimensional Normal Distribution for p≥3." Indiana University Technical Report.

Using Personal Computer Spreadsheets in Statistical Planning and Analysis

Carl T. Russell
US Army Operational Test and Evaluation Agency
Falls Church, Virginia

ABSTRACT. Personal computer spreadsheets provide an easy-to-use tool for performing many statistical computations. This paper describes examples of such computations. The first shows how the standard approximations for binomial sample sizing from Natrella can be implemented in a spreadsheet to produce flexible automatic tables. A second series of examples examines a variety of exact calculations involving binomial coefficients. Other spreadsheet applications are briefly discussed. These examples show that spreadsheets serve as alternatives or supplements to published tables or traditional programming languages for many statistical problems.

1. INTRODUCTION. This is a simple paper. Its thesis is that commercial microcomputer spreadsheet software is the first place one should look for assistance with many routine statistical computations. It was motivated by personal experience developing tabular displays, especially specialized tabular displays of discrete probability distributions. This experience showed that commercial microcomputer spreadsheet software (hereafter referred to as "spreadsheets") could be used quickly to implement versions of such probability tables. Little effort produces spreadsheet templates which can duplicate voluminous standard tables. More important, spreadsheets can produce custom interactive tables which provide quicker, more flexible and more accurate answers than standard tables. Most of the paper is devoted to examples of such probability tabulations, starting with an automated version of some standard binomial approximations and meandering through several exact calculations involving binomial coefficients. Other actual and potential applications are discussed briefly.

Some familiarity with spreadsheet software is required to appreciate this paper. Most important is realizing how easily formulas can be promulgated throughout an automated spreadsheet. Once appropriate combinations of absolute and relative references are devised, only a few formulas need be entered to generate a large, flexible table from a few input parameters. Moreover, only the relevant portion of the table needs to be examined, fine tuned, and printed.

2. ROUTINE BINOMIAL SAMPLE SIZING Á LA NATRELLA. Natrella's *Experimental Statistics*† is used widely in the Army for binomial sample sizing. The approach involves look-up from several tables based on an arcsine transformation. The theoretical basis for the approximations used by Natrella is that if X successes are observed in N Bernoulli trials with success probability p and f: $x \to f(x)$ is an appropriate arcsine transformation, then $Y=f(X)$ has approximately normal distribution with $\mu=\arcsin(\sqrt{p})$ and $\sigma^2=1/(4N)$; that is, the variance of Y does not depend on p. Natrella discusses four possibilities depending on whether there are one or two populations, and on whether one- or two-sided hypotheses are appropriate. That is, there are one- and two-sided hypotheses in each of two cases: one population compared against a standard (success probability p_0) and two populations compared against each other (success probabilities p_1 and p_2). In the one-population case,

$2\sqrt{N}(Y-p_0)$ is approximately normal with $\mu=2\sqrt{N}[\arcsin\sqrt{p} - \arcsin\sqrt{p_0}]$ and $\sigma^2=1$,

and in the two-population case,

$2\sqrt{N}(Y_1-Y_2)$ is approximately normal with $\mu=2\sqrt{N}[\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2}]$ and $\sigma^2=2$.

Writing down expressions for type I error (α) and type II error (β) and solving for N in terms of the percentiles of the standard normal distribution, $z_\kappa=\Phi^{-1}(\kappa)$, gives the formulas used

† National Bureau of Standards Handbook 91, 1963, US Government Printing Office, sections 8-1.4, 8-1.5, 8-2.4. This handbook was originally developed for limited distribution as US Army Ordnance Pamphlets ORDP 20-110 through 20-114. It is now supplemented and to a large extent replaced by DARCOM PAMPHLET No. 706-103, December 1983, which discusses binomial sample sizing in section 8-3.

	One-Sided	Two-Sided
One Population	$H_{11}: p \le p_0$ vs $K_{11}: p > p_0$ $N = \frac{(z_{1-\alpha}+z_{1-\beta})^2}{4(\arcsin\sqrt{p}-\arcsin\sqrt{p_0})^2}$	$H_{12}: p = p_0$ vs $K_{12}: p \ne p_0$ $N = \frac{(z_{1-\alpha/2}+z_{1-\beta})^2}{4(\arcsin\sqrt{p}-\arcsin\sqrt{p_0})^2}$
Two Populations	$H_{21}: p_1 \le p_2$ vs $K_{21}: p_1 > p_2$ $N = \frac{(z_{1-\alpha}+z_{1-\beta})^2}{2(\arcsin\sqrt{p_1}-\arcsin\sqrt{p_2})^2}$	$H_{22}: p_1 = p_2$ vs $K_{22}: p_1 \ne p_2$ $N = \frac{(z_{1-\alpha/2}+z_{1-\beta})^2}{2(\arcsin\sqrt{p_1}-\arcsin\sqrt{p_2})^2}$

Figure 1. Binomial Sample Sizing Formulas Used by Natrella.

in Natrella's tables (see Figure 1). Implementing these formulas in a spreadsheet is easy. Each numerator in the sample size formula, $\delta^2=(z_{1-\alpha}+z_{1-\beta})^2$ or $\delta^2=(z_{1-\alpha/2}+z_{1-\beta})^2$, depends only on α and β, and each denominator, $d^2=4(\arcsin\sqrt{p}-\arcsin\sqrt{p_0})^2$ or $d^2=4(\arcsin\sqrt{p_1}-\arcsin\sqrt{p_2})^2$, depends only on the parameters specified by simple null and alternative hypotheses. Interactive spreadsheet tables for binomial sample sizing can be built by providing a data entry area for choosing α, entering the difference in parameters to be detected ($\Delta=p-p_0$ or $\Delta=p_1-p_2$), and specifying the range of parameters to be examined. Figure 2 assumes that such a data entry area has specified—in addition to Δ and α—an initial probability value p_{00} and a value λ to be used to increment p_{00} for a fixed number of lines. Both α and β require Φ^{-1}, which is not available as a standard spreadsheet function. A macro could probably be written to compute Φ^{-1}, but a simpler approach is to limit choices for α to a few values and let "confidence levels" $1-\beta$ vary across fixed standard values. That was done in the example of Figure 2 and the actual spreadsheet implementation in Figure 3. The original motivation for the spreadsheet in Figure 3 was to help an evaluator assess a resource requirement for an operational test. An analyst was arguing on the basis of Natrella's formulas that in order to have 80% confidence of detecting a 10 probability point difference in kill probabilities between two missile systems, about 115 missiles of each type would be required—versus the 100 missiles of each type which were available. This claim was based on two "assumptions": an assumption that the difference to be detected was between $p_2=0.85$ and $p_1=0.95$ and an assumption that 10% to 15% of any firings would be "no tested." The sensitivity of binomial sample sizes to assumptions about the underlying probabilities was not clear to the evaluator. Once the *table* in Figure 3 was produced, the evaluator could see that since the underlying probabilities could just as well be $p_2=0.75$ and $p_1=0.85$ or even $p_2=0.65$ and $p_1=0.75$, obtaining a few more expensive missiles was not something to fall on his sword over. The simple capability to produce a *full table* instead of a *few numbers* provided a convincing test planning tool.

			δ_1	...	δ_5
			$1-\beta_1$	...	$1-\beta_5$
p_2	p_1	d	Required Sample Size		
p_{00}	$p_2+\Delta$	$d(p_1,p_2)$	$2\delta_1^2/d^2$	...	$2\delta_5^2/d^2$
$p_{00}+\lambda$	$p_2+\Delta$	$d(p_1,p_2)$	$2\delta_1^2/d^2$	...	$2\delta_5^2/d^2$
$(p_{00}+\lambda)+\lambda$	$p_2+\Delta$	$d(p_1,p_2)$	$2\delta_1^2/d^2$	...	$2\delta_5^2/d^2$
⋮	⋮	⋮	⋮	⋮	⋮

Figure 2. Example of Spreadsheet Template for Natrella's Sample Size Formula—Two-Sided Test Between Two Observed Proportions.

3. IMPLEMENTING EXACT BINOMIAL TABLES. At one time or another, nearly every applied statistician has attempted to program exact calculations for probabilities based on binomial coefficients. Using FORTRAN or BASIC, potential underflow and overflow must be carefully considered to avoid silly answers. Using a spreadsheet, calculations more accurate than standard tables can be obtained with very little care. Figure 4 shows the key binomial

TWO-SIDED HYPOTHESIS TEST FOR DIFFERENCE BETWEEN TWO OBSERVED PROPORTIONS
SAMPLE SIZE REQUIREMENTS FROM NATRELLA 8-2.4.1, PAGES 8-18 & 8-19

Sample Size=2*(delta)**2/d**2
where delta=z[1-alpha/2]+z[1-beta],
d=2*(arcsin(sqrt(P'))-arcsin(sqrt(P''))),
and z[k]=kth percentile of the standard normal.

Enter: Starting P'is 0.450 Diff for P''=P'+Diff is 0.100
Increment for P' is 0.010
Significance Level alpha (use: .01, .05, .1 or .2) is 0.10
alpha used is 0.10 z[1-alpha/2] is 1.65

	Delta=	3.97	3.29	2.93	2.49	2.17	1.90	1.65
Conf	(1-beta)=	0.99	0.95	0.90	0.80	0.70	0.60	0.50

P'	P''	d	Sample Size Required to Obtain Prescribed Confidence That P' Differs From P'' at Significance Level alpha						
0.450	0.550	0.200	786	540	427	309	235	180	135
0.460	0.560	0.200	786	540	427	309	235	180	135
0.470	0.570	0.200	785	539	427	308	235	180	135
0.480	0.580	0.201	783	538	426	308	234	179	135
0.490	0.590	0.201	781	536	425	307	233	179	135
0.500	0.600	0.201	778	534	423	306	233	178	134
0.510	0.610	0.202	775	532	421	304	232	177	133
0.520	0.620	0.202	771	529	419	303	230	176	133
0.530	0.630	0.203	766	526	416	301	229	175	132
0.540	0.640	0.204	760	522	413	299	227	174	131
0.550	0.650	0.205	754	518	410	296	225	173	130
0.560	0.660	0.205	748	513	406	294	223	171	129
0.570	0.670	0.206	740	508	402	291	221	170	127
0.580	0.680	0.208	732	503	398	288	219	168	126
0.590	0.690	0.209	724	497	393	284	216	166	125
0.600	0.700	0.210	715	491	388	281	214	164	123
0.610	0.710	0.212	705	484	383	277	211	161	121
0.620	0.720	0.213	694	477	377	273	207	159	120
0.630	0.730	0.215	683	469	371	268	204	156	118
0.640	0.740	0.217	671	461	365	264	201	154	116
0.650	0.750	0.219	659	452	358	259	197	151	113
0.660	0.760	0.221	646	443	351	253	193	148	111
0.670	0.770	0.224	632	434	343	248	189	145	109
0.680	0.780	0.226	617	424	336	242	185	141	106
0.690	0.790	0.229	602	414	327	237	180	138	104
0.700	0.800	0.232	587	403	319	230	175	134	101
0.710	0.810	0.235	570	392	310	224	170	131	98
0.720	0.820	0.239	553	380	301	217	165	127	95
0.730	0.830	0.243	535	368	291	210	160	123	92
0.740	0.840	0.247	517	355	281	203	155	118	89
0.750	0.850	0.252	498	342	271	196	149	114	86
0.760	0.860	0.257	478	328	260	188	143	110	82
0.770	0.870	0.263	458	314	249	180	137	105	79
0.780	0.880	0.269	437	300	237	172	131	100	75
0.790	0.890	0.276	415	285	226	163	124	95	72
0.800	0.900	0.284	392	269	213	154	117	90	68
0.810	0.910	0.293	369	253	201	145	110	85	64
0.820	0.920	0.303	345	237	187	135	103	79	60
0.830	0.930	0.314	319	219	174	126	96	73	55
0.840	0.940	0.328	293	202	160	115	88	67	51
0.850	0.950	0.344	266	183	145	105	80	61	46
0.860	0.960	0.364	238	164	130	94	71	55	41
0.870	0.970	0.390	208	143	113	82	63	48	36
0.880	0.980	0.424	176	121	96	69	53	41	31
0.890	0.990	0.476	140	96	76	55	42	32	24

Figure 3. Example Based on the Natrella Formulas.
(Printout of an Enable Spreadsheet on a Zenith 248.)

Key Relationship: $\binom{N}{k+1} \Big/ \binom{N}{k} = \frac{N-k}{k+1}$

	A	B	C	D	E	J	K
1	*N*	*Initial p*	*Delta p*				
2			$\binom{N}{K}$	B1	D2+C1	D2	E2
3	k	N-k		1-D2	1-E2	D3	E3
4							
5	0	A1-A5	1	$C5*D$2^$A5*D$3^B5	...	D5	E5
6	A5+1	A1-A6	C5*B5/A6	$C6*D$2^$A6*D$3^B6	...	J5+D6	K5+E6
7	A6+1	A1-A7	C6*B6/A7	$C7*D$2^$A7*D$3^B7	...	J6+D7	K6+E7
:	:	:	:	:	:	:	:

Figure 4. Spreadsheet Template for Exact Binomial Tables.

relationship and formulas which generate tables of both individual and cumulative binomial probabilities. *Script type* in Figure 4 indicates a data entry area while gray type indicates formulas obtained by "copying," which can be extended at will. The notation in Figure 4 is the standard from Lotus 1-2-3 with columns labeled by letters and rows by numerals and with "$" indicating an absolute rather than the default relative reference. In Figure 4, columns D-H contain individual probabilities while columns J-M contain the cumulative probabilities. Column C—which contains the binomial coefficients—needs to be calculated but is not of direct interest; its display would normally be suppressed. Likewise, formatting or logic tricks can be used to suppress printing many values very close to zero or one, as in Figure 5.

Figure 5 shows a portion of a large table of binomial probabilities generated via a template similar to that of Figure 4. The only substantial difference is that Figure 5 displays nine values for p vice five in Figure 4, and the p-values in Figure 5 are controlled by a center value and a delta in both directions vice a starting value and a delta in Figure 4. The fact that all cumulative columns end in 1.0000000 confirms substantial numerical accuracy. Underlying spreadsheet calculations are typically performed to 14 significant figure accuracy, so multiplication of the very large binomial coefficients with the very small products of success and failure probabilities is accurate to nearly 14 significant figures, and only very tiny probabilities are lost to underflow when cumulated. Since standard tables typically display only 7-place accuracy—already more than needed for practical purposes—accuracy of spreadsheet calculations presents no problem. Memory and computing time is a greater concern. On the standard Apple Macintosh SE where Figure 5 was calculated and printed (using Microsoft Excel), loading or recalculating the spreadsheet takes several minutes, the spreadsheet loaded into Excel takes approximately 760 kilobytes of memory, and storage of the spreadsheet takes more than 500 kilobytes on disk. (A similar spreadsheet implemented in Enable on a Zenith 248 with 640 kilobytes of RAM runs out of memory when N is slightly larger than 100.)

4. RETHINKING TABULATION OF DISCRETE PROBABILITIES. For practical purposes, the spreadsheet template in Figure 4 (implemented in Figure 5), replaces all standard binomial tables. Templates for other discrete distributions requiring binomial coefficients are also easy to implement, both for standard distributions such as the hypergeometric distribution and for more unusual distributions such as that tabulated in Figure 6. Unlike previous tables in this paper, Figure 6 represents a rethinking of probability tabulation rather than a straightforward translation of traditional tables into an automated spreadsheet. It shows the screen image of an Excel spreadsheet on a Macintosh, formatted for ease of interactive sample

Parameters Entered: N= 367 MidP= 0.95 DeltaP= 0.01

BINOMIAL DISTRIBUTION: PROBABILITY OF k SUCCESSES IN N TRIALS

k	N-k	Binomial Coefficient	Success Probability P [Failure Probability Q] 0.910 [0.090]	0.920 [0.080]	0.930 [0.070]	0.940 [0.060]	0.950 [0.050]	0.960 [0.040]	0.970 [0.030]	0.980 [0.020]	0.990 [0.010]
306	61	2.77762528E+70									
307	60	5.51906002E+69									
308	59	1.07514158E+69	0.00001								
309	58	2.05285829E+68	0.00001								
310	57	3.84083352E+67	0.00002								
311	56	7.03846979E+66	0.00004								
312	55	1.26348458E+66	0.00006								
313	54	2.22019814E+65	0.00011	0.00001							
314	53	3.81817514E+64	0.00020	0.00001							
315	52	6.42423119E+63	0.00034	0.00002							
316	51	1.05715197E+63	0.00056	0.00004							
317	50	1.70078077E+62	0.00091	0.00008							
318	49	2.6741836E+61	0.00145	0.00015	0.00001						
319	48	4.10788014E+60	0.00225	0.00026	0.00001						
320	47	6.16152021E+59	0.00341	0.00044	0.00003						
321	46	9.02154048E+58	0.00504	0.00075	0.00005						
322	45	1.2887915E+58	0.00728	0.00123	0.00010						
323	44	1.79552985E+57	0.01026	0.00197	0.00015	0.00001					
324	43	2.43837401E+56	0.01409	0.00307	0.00033	0.00001					
325	42	3.22615638E+55	0.01885	0.00467	0.00058	0.00003					
326	41	4.15638779E+54	0.02456	0.00692	0.00098	0.00006					
327	40	5.21138561E+53	0.03113	0.00998	0.00164	0.00011					
328	39	6.3553483E+52	0.03838	0.01403	0.00266	0.00022	0.00001				
329	38	7.53380556E+51	0.04601	0.01908	0.00418	0.00040	0.00001				
330	37	8.67516458E+50	0.05357	0.02527	0.00640	0.00073	0.00003				
331	36	9.68731383E+49	0.06055	0.03248	0.00951	0.00127	0.00006				
332	35	1.05151586E+49	0.06638	0.04051	0.01370	0.00217	0.00012				
333	34	1.10519686E+48	0.07055	0.04885	0.01912	0.00357	0.00025				
334	33	1.1250508E+47	0.07261	0.05732	0.02586	0.00568	0.00048	0.00001			
335	32	1.10825889E+46	0.07232	0.06493	0.03385	0.00878	0.00088	0.00002			
336	31	1.05548476E+45	0.06964	0.07111	0.04283	0.01308	0.00161	0.00005			
337	30	9.70820695E+43	0.06478	0.07522	0.05235	0.01887	0.00281	0.00012			
338	29	8.6176383E+42	0.05813	0.07679	0.06173	0.02624	0.00474	0.00025			
339	28	7.37202182E+41	0.05028	0.07554	0.07015	0.03517	0.00771	0.00052	0.00001		
340	27	6.07107679E+40	0.04187	0.07154	0.07676	0.04538	0.01206	0.00103	0.00001		
341	26	4.80701085E+39	0.03352	0.06514	0.08074	0.05628	0.01815	0.00195	0.00004		
342	25	3.65445277E+38	0.02577	0.05695	0.08155	0.06704	0.02621	0.00356	0.00009		
343	24	2.66358531E+37	0.01899	0.04774	0.07897	0.07655	0.03630	0.00622	0.00022		
344	23	1.85832231E+36	0.01338	0.03830	0.07320	0.08367	0.04812	0.01042	0.00049		
345	22	1.23868154E+35	0.00903	0.02936	0.06483	0.08738	0.06096	0.01667	0.00106		
346	21	7.87728147E+33	0.00580	0.02147	0.05477	0.08705	0.07364	0.02544	0.00218	0.00002	
347	20	4.76723986E+32	0.00355	0.01494	0.04404	0.08254	0.08468	0.03695	0.00427	0.00005	
348	19	2.73878785E+31	0.00206	0.00988	0.03362	0.07432	0.09246	0.05096	0.00794	0.00013	
349	18	1.49157505E+30	0.00114	0.00618	0.02432	0.06339	0.09564	0.06658	0.01387	0.00034	
350	17	7.67095739E+28	0.00059	0.00366	0.01662	0.05107	0.09345	0.08219	0.02323	0.00085	
351	16	3.71527851E+27	0.00029	0.00204	0.01069	0.03875	0.08600	0.09553	0.03638	0.00203	
352	15	1.58876296E+26	0.00013	0.00106	0.00646	0.02760	0.07427	0.10422	0.05346	0.00451	
353	14	7.17604657E+24	0.00006	0.00052	0.00365	0.01837	0.05997	0.10628	0.07345	0.00940	0.00002
354	13	2.83798452E+23	0.00002	0.00024	0.00192	0.01138	0.04506	0.10086	0.09392	0.01822	0.00008
355	12	1.33826194E+22	0.00001	0.00010	0.00093	0.00653	0.03135	0.08866	0.11121	0.03268	0.00028
356	11	3.50313012E+20		0.00004	0.00042	0.00345	0.02008	0.07173	0.12121	0.05398	0.00088
357	10	1.07939584E+19		0.00001	0.00017	0.00168	0.01175	0.05304	0.12075	0.08151	0.00298
358	9	3.01507217E+17			0.00006	0.00073	0.00624	0.03555	0.10906	0.11156	0.00825
359	8	7.55867674E+15			0.00002	0.00029	0.00297	0.02138	0.08840	0.13704	0.02048
360	7	1.67970594E+14			0.00001	0.00010	0.00125	0.01141	0.06352	0.14922	0.04507
361	6	3257047534741				0.00003	0.00046	0.00531	0.03002	0.14178	0.08652
362	5	53984213283				0.00001	0.00015	0.00211	0.02134	0.11515	0.14197
363	4	743584205					0.00006	0.00070	0.00951	0.07772	0.19380
364	3	8171255					0.00001	0.00018	0.00336	0.04185	0.21062
365	2	67161						0.00004	0.00080	0.01685	0.17138
366	1	367							0.00016	0.00451	0.08271
367	0	1							0.00001	0.00060	0.02501

CUM. BINOMIAL DISTRIBUTION: PROB. OF AT LEAST k SUCCESSES IN N TRIALS

k	N-k	Success Probability P [Failure Probability Q] 0.910 [0.090]	0.920 [0.080]	0.930 [0.070]	0.940 [0.060]	0.950 [0.050]	0.960 [0.040]	0.970 [0.030]	0.980 [0.020]	0.990 [0.010]
306	61	0.0000025								
307	60	0.0000052	0.0000001							
308	59	0.0000104	0.0000003							
309	58	0.0000204	0.0000006							
310	57	0.0000394	0.0000013							
311	56	0.0000747	0.0000027							
312	55	0.0001386	0.0000057	0.0000001						
313	54	0.0002522	0.0000117	0.0000002						
314	53	0.0004490	0.0000236	0.0000005						
315	52	0.0007858	0.0000466	0.0000012						
316	51	0.0013451	0.0000901	0.0000027						
317	50	0.0022548	0.0001706	0.0000058	0.0000001					
318	49	0.0037011	0.0003162	0.0000123	0.0000002					
319	48	0.0059472	0.0005734	0.0000256	0.0000004					
320	47	0.0093538	0.0010170	0.0000522	0.0000010					
321	46	0.0143873	0.0017640	0.0001038	0.0000023					
322	45	0.0216822	0.0029912	0.0002017	0.0000053					
323	44	0.0319442	0.0049573	0.0003831	0.0000118	0.0000001				
324	43	0.0460351	0.0080280	0.0007103	0.0000256	0.0000003				
325	42	0.0648857	0.0127000	0.0012854	0.0000543	0.0000007				
326	41	0.0894414	0.0196221	0.0022680	0.0001122	0.0000017				
327	40	0.1205721	0.0296031	0.0039086	0.0002258	0.0000042				
328	39	0.1589582	0.0436004	0.0065665	0.0004432	0.0000098				
329	38	0.2049671	0.0626827	0.0107506	0.0008467	0.0000227	0.0000001			
330	37	0.2585357	0.0878519	0.0171523	0.0015746	0.0000508	0.0000004			
331	36	0.3190813	0.1204353	0.0266501	0.0028483	0.0001108	0.0000010			
332	35	0.3854626	0.1609417	0.0403548	0.0050148	0.0002336	0.0000026			
333	34	0.4560080	0.2098022	0.0594796	0.0085806	0.0004793	0.0000067			
334	33	0.5286186	0.2672182	0.0853446	0.0142674	0.0009545	0.0000168			
335	32	0.6009402	0.3321479	0.1191952	0.0230437	0.0018438	0.0000401	0.0000001		
336	31	0.6705832	0.4032613	0.1620265	0.0361385	0.0034532	0.0000938	0.0000003		
337	30	0.7353583	0.4784896	0.2143720	0.0550101	0.0062960	0.0002126	0.0000010		
338	29	0.7934896	0.5552758	0.2760079	0.0812516	0.0110085	0.0004654	0.0000030		
339	28	0.8437714	0.6308163	0.3462518	0.1164208	0.0187194	0.0009844	0.0000086		
340	27	0.8856389	0.7023576	0.4230083	0.1617962	0.0307831	0.0020103	0.0000233		
341	26	0.9191582	0.7675000	0.5037522	0.2180829	0.0489318	0.0039588	0.0000610		
342	25	0.9449248	0.8244520	0.5853055	0.2851222	0.0751465	0.0075166	0.0001538	0.0000002	
343	24	0.9639131	0.8721866	0.6642774	0.3616734	0.1114487	0.0137380	0.0003718	0.0000006	
344	23	0.9773080	0.9104882	0.7374773	0.4453457	0.1585726	0.0241575	0.0008644	0.0000021	
345	22	0.9863371	0.9388528	0.8023114	0.5327367	0.2205282	0.0408271	0.0019260	0.0000070	
346	21	0.9921420	0.9613239	0.8570805	0.6197908	0.2941662	0.0662651	0.0041084	0.0000222	
347	20	0.9958040	0.9762671	0.9011167	0.7023294	0.3788436	0.1032125	0.0083790	0.0000673	
348	19	0.9977581	0.9861433	0.9347405	0.7766457	0.4713053	0.1541744	0.0163146	0.0001944	
349	18	0.9988043	0.9923265	0.9580602	0.8400310	0.5688461	0.2207606	0.0332839	0.0005333	0.0000001
350	17	0.9994851	0.9959835	0.9756770	0.8911014	0.6604008	0.3029[illegible]	0.0535119	0.0013673	0.0000003
351	16	0.9997745	0.9980203	0.9863694	0.9298524	0.7464008	0.3864805	0.0898875	0.0034142	0.0000014
352	15	0.9999074	0.9990650	0.9928265	0.9574486	0.8206731	0.5028865	0.1433436	0.0079286	0.0000063
353	14	0.9999646	0.9996053	0.9964719	0.9758197	0.8806361	0.6088831	0.2168008	0.0173282	0.0000259
354	13	0.9999874	0.9998420	0.9983872	0.9872021	0.9258642	0.7098634	0.3107258	0.0355432	0.0001078
355	12	0.9999959	0.9999416	0.9993191	0.9937324	0.9570471	0.7985244	0.4219361	0.0682277	0.0004011
356	11	0.9999988	0.9999802	0.9997364	0.9971808	0.9771255	0.8702502	0.5431430	0.1222122	0.0013796
357	10	0.9999997	0.9999939	0.9999072	0.9988456	0.9888801	0.9232911	0.6638073	0.2037183	0.0043646
358	9	0.9999999	0.9999983	0.9999706	0.9995741	0.9951186	0.9586482	0.7728584	0.3152788	0.0126193
359	8	1.0000000	0.9999996	0.9999917	0.9998602	0.9980601	0.9802435	0.8613617	0.4523167	0.0331054
360	7		0.9999999	0.9999980	0.9999580	0.9993448	0.9916538	0.9248811	0.6015378	0.0731780
361	6		1.0000000	0.9999996	0.9999901	0.9998070	0.9968639	0.9647053	0.7433186	0.1647006
362	5			0.9999999	0.9999980	0.9999526	0.9990762	0.9860476	0.8584668	0.3066740
363	4			1.0000000	0.9999997	0.9999907	0.9997744	0.9955527	0.9361835	0.5002742
364	3				1.0000000	0.9999986	0.9999568	0.9988298	0.9780311	0.7108942
365	2					0.9999998	0.9999940	0.9998275	0.9948847	0.8822754
366	1					1.0000000	0.9999997	0.9999860	0.9993875	0.9749888
367	0						1.0000000	1.0000000	1.0000000	1.0000000

Figure 5. Example of an Exact Binomial Table.
(Printout of an Excel Spreadsheet on a Macintosh SE.)

sizing or inference. Only two values for p are shown since only two are typically needed for sample sizing or determining confidence intervals. Standard tabulated values are supplemented by several useful arithmetic results, and the data entry area is arranged so that it always remains on screen. The underlying distribution comes from a series of Bernoulli trials where R rounds are fired at T targets (T≤R) until either all rounds are expended or all targets are killed. When T=R the distribution is binomial, and when T<<R it approximates the negative binomial distribution. The R+1 possible outcomes (indexed by M) are as follows:

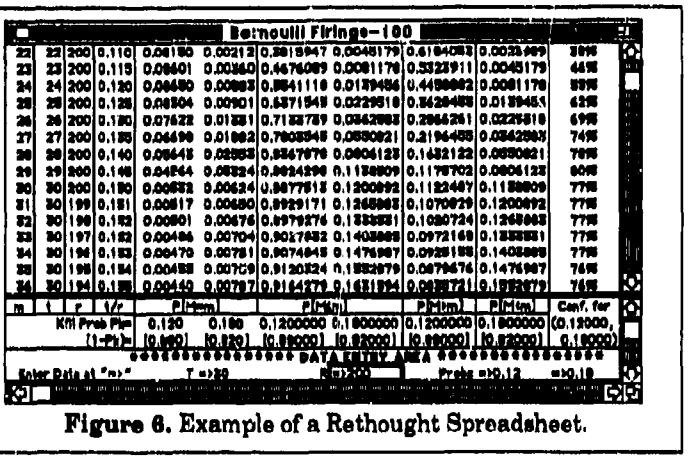


Figure 6. Example of a Rethought Spreadsheet.

$$M=0: \quad \text{Prob}\{T=0 \text{ targets killed with } R=R \text{ rounds}\} = \binom{R}{0}(1-p)^R$$

$$M=1: \quad \text{Prob}\{T=1 \text{ targets killed with } R=R \text{ rounds}\} = \binom{R}{1}p(1-p)^{R-1}$$

$$\cdots \qquad \cdots$$

$$M=t: \quad \text{Prob}\{T=t \text{ targets killed with } R=R \text{ rounds}\} = \binom{R}{t}p^t(1-p)^{R-t}$$

$$\cdots \qquad \cdots$$

$$M=T-1: \quad \text{Prob}\{T=T-1 \text{ targets killed with } R=R \text{ rounds}\} = \binom{R}{T-1}p^{T-1}(1-p)^{R-T+1}$$

$$M=T: \quad \text{Prob}\{T=T \text{ targets killed with } R=R \text{ rounds}\} = \binom{R}{T}p^T(1-p)^{R-T} = \binom{R-1}{T-1}p^T(1-p)^{R-T}$$

$$M=T+1: \quad \text{Prob}\{T=T \text{ targets killed with } R=R-1 \text{ rounds}\} = \binom{R-2}{T-1}p^T(1-p)^{R-T-1}$$

$$\cdots \qquad \cdots$$

$$M=T+k: \quad \text{Prob}\{T=T \text{ targets killed with } R=R-k \text{ rounds}\} = \binom{R-k-1}{T-k}p^{T-k}(1-p)^{R-T-k}$$

$$\cdots \qquad \cdots$$

$$M=R: \quad \text{Prob}\{T=T \text{ targets killed with } R=T \text{ rounds}\} = \binom{T-1}{T-1}p^T$$

Since the data entry area and column labels at the bottom of the screen do not scroll with the body of the table, parameters can be changed easily and the results observed immediately. From a sample sizing viewpoint, the screen in Figure 6 shows that with T=30 targets and R=200 rounds, observed kill proportions near 0.15 will produce 80% confidence intervals somewhat less than 0.1 in length. Once data are collected, the same spreadsheet can be used to determine confidence intervals. Figure 7 shows that observing 26 kills when T=30 and R=200 yields (0.1044,0.1603) as an 80% confidence interval. Alternatively, the same spreadsheet could be used to investigate

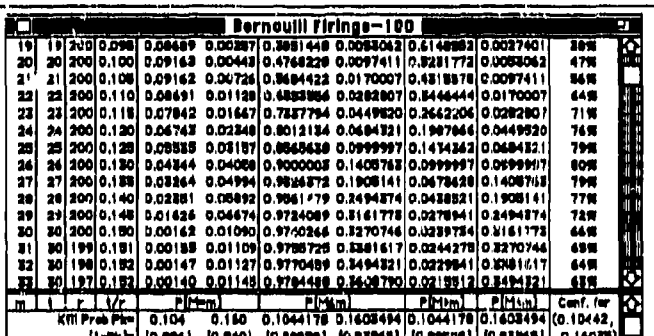


Figure 7. 80% Confidence Interval for T=26.

other sample sizes. Figure 8 shows that T~11 and R~74 provide the approximate sample size required to detect a 0.1 difference in kill probabilities with 80% confidence. Although less than 40 rows have been displayed in Figures 6-8, the spreadsheet was laid out with 100 rows for flexibility (more could be obtained by copying rows downward if necessary). Since this spreadsheet is much more compact and requires fewer demanding calculations than a full table like that of Figure 5, it recalculates much faster (less than 10 seconds). Thus the iterative fiddling required to obtain results such as those in Figures 7 and 8 is quite feasible.

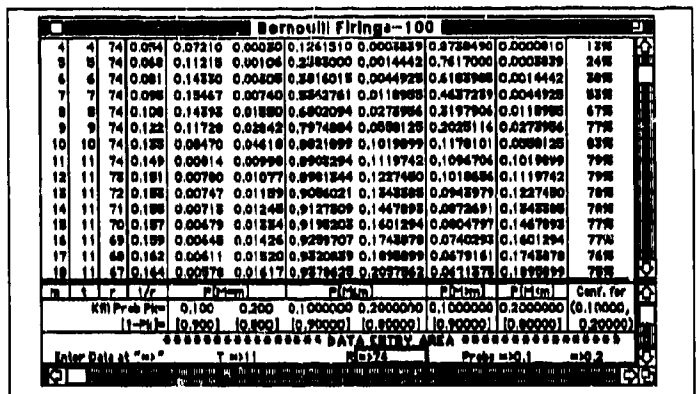


Figure 8. Sample Size Needed to Detect $\Delta p=0.10$.

5. OTHER ACTUAL AND POTENTIAL APPLICATIONS. Use of spreadsheets for statistical calculations is not limited to calculation of probability tables. In particular, spreadsheets are useful in conjunction with other programs which perform statistical analyses. Arithmetic operations are frequently required to understand, interpret, and present the results of analyses performed using standard statistical packages. Spreadsheets can reduce the manual labor involved with such operations without requiring specialized programming. For example, SAS Least Squares Means (LSM's) provide representations for various marginal means as if the underlying experimental design had been balanced. SAS can calculate LSM's for any effect in an underlying model, but cannot calculate LSM's for any effect not in a model. Simple but tedious arithmetic can be used to calculate internal values from margins for presentation. If more than one or two such calculations is to be done, writing a spreadsheet template to do them pays off. Similarly, a spreadsheet can provide a convenient way of translating back and forth between estimates obtained on a transformed variable and more easily understood corresponding estimates on the untransformed variable—for instance, translating results of an analysis of $\log(Y+0.02)$ back into statements about $(p-p_0)/p_0$, where $E_{\text{expectation}}[\log(Y+0.02)]=\log(p+0.02)$. Still another related application was suggested following presentation of this paper by a statistician who routinely uses spreadsheets in conjunction with other procedures to perform jackknifing. Finally, since spreadsheets read and write files consisting of tab delineated fields, automated exchange of data with other computer programs can be easy. Spreadsheet capabilities for editing and rearranging data make them a good preprocessor for specialized statistical packages like MacSpin, which have less flexible data entry capabilities. Additional capabilities of most modern spreadsheets include macro language capabilities, which make nonstandard formulas and calculations readily available, and integrated graphics capabilities, some of which are quite good. Every statistician having access to a microcomputer should understand the kinds of things spreadsheets can do to make life easier, both as stand-alone tools and as supplements to other tools. The convenient power of microcomputer spreadsheets provides computational tools which should be the first place one looks for assistance with routine statistical calculations.

Attendees: 33rd U.S. Army Design of Experiments Conference

Altekar, Maneesha
University of Delaware
7 Fairway Road
Apt. 1-C
Newark, DE 19711

Andriolo, Miquel
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
ATTN: SLCBR-SE-D

Baker, William E.
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

Bates, Carl B.
U. S. Army Concepts Analylsis Agency
8120 Woodmont Avenue
Bethesda, MD 20814-2797

Bissinger, Barney
Department of Mathematical Sciences
Pennsylvania State College
at Harrisburg
Box 14
Capitol College
Middletown, PA 17057

Bodt, Barry A.
Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

Brodeen, Ann E. M.
U. S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
ATTN: SLCBR-SE-D

Celmins, Aivars
U. S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

Crama, Yves
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Crosier, Ronald
U. S. Army Chemical Research, Development
and Engineering Center
Aberdeen Proving Ground, MD 21010-5423
ATTN: SMCCR-RSP-C

Dewald, Lee S., Sr.
Department of Mathematics
U. S. Military Academy
West Point, NY 10996

Ding, Yijun
Department of Mathematics Sciences
University of Delaware
Newark, DE 19716

Dressel, Francis
Department of the Army
U. S. Army Laboratory Command
Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

Dutoit, Eugene
U. S. Army Infantry School
Fort Benning, GA 31907

Essenwanger, Oskar M.
U. S. Army Missile Command
Reserach Directorate, RD&E Center
Redstone Arsenal, AL 35898-52418
ATTN: SLCBR-SE-D

Federer, Walter T.
Mathematical Sciences Institute
337 Warren Hall
Cornell University
Ithaca, NY 14850

Fernandez, Joel H.
U.S. Army Material Test and Evaluation Directorate
White Sanda Missile Range, NM 88002-5175

Ganju, Jitendra
Department of Mathematical Sciences
University of Delaware
Newark, DE 19713

Gehle, Harold
Army Management Engineering
Training Activity
Rock Island, IL 61299-7040

Gorman, Robert
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Grubbs, Frank E.
4109 Webster Road
Havre De Grace, MD 21078

Grynovicki, Jock O.
Human Engineering Laboratory
U. S. Army Laboratory Command
Aberdeen Proving Ground, MD 21005-5001
ATTN: SLCHE-HU/EDP/Mr. Jock O. Grynovicki

Hoerl, Arthur E.
Department of Mathematics
University of Delaware
Newark, DE 19716

Hunter, Charles J.
Department of National Defence
Air Transport Group Headquarters
Operation Research Adv.
Astra, Ontario K0K 1B0
CANADA

Hunter, J. Stuart (Retired)
503 Lake Drive
Princeton, NJ 08540

Jacqmein, William M.
U.S.A. Operational Test & Evaluation Agency
5600 Colulmbia Pike
Falls Church, VA 22041-5115

Jackson, William
U.S. Army Tank - Automotive Command
AMSTA-RSA
Warren, MI 48397-5000

Janssens, Radford
Department of Mathematical Sciences
50 W. Delaware
University of Delaware
Newark, DE 19716

Khan, Anqeel A.
U. S. Army Concepts Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20814-2797

Kolb, Rickey A.
Department of Mathematics
U. S. Military Academy
West Point, NY 10996

LaRiccia, Vincent N.
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Launer, Robert
Department of the Army
U. S. Army Laboratory Command
Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

Lien, John N.
U. S. Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
ATTN: SLCBP-SE-D

Masciantonio, Mario
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Mehrotra, Devan
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Mitchell, Toby J.
Oak Ridge National Laboratory
Building 9207A, MD - 3
Box Y
Oak Ridge, TN 39831

Moss, Linda L. Crawford
U. S. Army Ballistic
Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
ATTN: SLCBR-SE-D

Nair, Srikantan S.
U. S. Army OTEA
5600 Columbia Pike
Falls Church, VA 22041

Parzen, Emanuel
Department of Statistics
Texas A&M University
College Station, TX 77840-3143

Pippin, Kathryn A.
Department of Correction
80 Monrovia Avenue
Smyrna, DE 18977

Planner John Stanley
Department of Correction
80 Monrovia Avenue
Smyrna, De 18977

Porter, Katherine
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Post, William
Department of Correction
80 Monrovia Avenue
Smyrna, DE 19977

Prabhu, Narahari U.
Mathematical Sciences Institute
Cornell University
Ithaca, NY 14853

Roediger, Paul A.
AMSMC-QAH-A(D) Pic
Picatinny Arsenal, NJ 07806-5000

Russ, Edward W.
Worcester Polytechnic Institute
Worcester, MA 01609

Sacher, Richard S.
Academic Computing Services
University of Delaware
Newark, DE 19716

Sams, Michelle R.
1120 Rainbow Drive
Las Cruces, NM 88005

Schuenemeyer, Jack
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Spears, Linda
U. S. Army Cold Regions Test Center
Alaska, APO Seattle 98733
ATTN: STECR-MT-A

Stakgold, Ivar
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Stark, Robert M.
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Sturdivan, Larry
U. S. Chemical Research, Development
and Engineering Center
Aberdeen Proving Ground, MD 21010-5423
ATTN: SMCCR-RSP-C

Taylor, Malcolm S.
U. S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
ATTN: SLCBR-SE-D

Testerman, Deloris
Analysis & Test Services
Yuma Proving Ground, AZ
ATTN: STEYP-MT-ES

Thomas, Jerry
U. S. Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

Thompson, Andrew A.
AMSAA
AMXSY-CA APG 21005-5071

Thrasher, Paul H.
U. S. Plans & Quality Assurance Directorate
White Sands Missile Range, NM 88002
ATTN: STEWS-PL-Q

Tingey, Henry B.
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Tung, Sarah T. Y.
Academic Computer Services
University of Delaware
Newark, DE 19716

Umholtz, Robert L.
U. S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
ATTN: SLCBR-SE-D

Webb, Dave
U. S. Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
ATTN: SLCBR-SE-D

Weinberger, Marcus A.
Operational Research and Analysis Establishment
Department of National Defence
Ottawa, Ontario
K1A OK2 Canada

Willard, Daniel
Department of the Army
Office of the Deputy Secretary of the Army (OR)
Room 2E660
Pentagon, Washington, DC 20310-0102

Winner, Wendy A.
U. S. Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
ATTN: SLCBR-SE-D

Wolff, Weston C.
White Sands Missile Range, NM 88002-5134
ATTN: STEWS-NR-CF

Womacki, Franklin
U. S. Army Concepts Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20814

Woods, Anthony
U.S. Army Troop Support Command
4300 Goodfellow Boulevard
St. Louis, MO 63120-1798

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188
Exp. Date Jun 30, 1986

1a. REPORT SECURITY CLASSIFICATION	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ARO Report 88-2	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Army Research Office 6b. OFFICE SYMBOL (If applicable) SLCRO-ARO	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Research Triangle Park, NC 27709-2211	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AMSC on behalf of ASO (ADA) 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) Proceedings of the Thirty-Third Conference on the Design of Experiments in Army Research, Development, and Testing	
12. PERSONAL AUTHOR(S)	
13a. TYPE OF REPORT Technical 13b. TIME COVERED FROM 88 Jan TO 89 Feb	14. DATE OF REPORT (Year, Month, Day) 1989, May 15. PAGE COUNT 291
16. SUPPLEMENTARY NOTATION	
17. COSATI CODES FIELD / GROUP / SUB-GROUP	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This is a technical report of the Thirty-Third Conference on the Design of Experiments in Army Research, Development, and Testing. It contains most of the papers presented at this meeting. These articles treat various Army statistical and design problems.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Francis G. Dressel	22b TELEPHONE (Include Area Code) (919) 549-0641 22c. OFFICE SYMBOL SLCRO-MA

DD FORM 1473, 84 MAR

83 APR edition may be used until exhausted
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED